# Graphical Models and Knowledge Discovery + Algorithms and Approximation Methods for Imprecise Probabilities

Two Lectures on IPs by

Cassio Polpo de Campos, Alessandro Antonucci, and Francesca Mangili

{cassio,alessandro,francesca}@idsia.ch

Istituto "Dalle Molle" di Studi sull'Intelligenza Artificiale - Lugano (Switzerland)

Sixth SIPTA Summer School, Montpellier, July 21st and 22nd, 2014

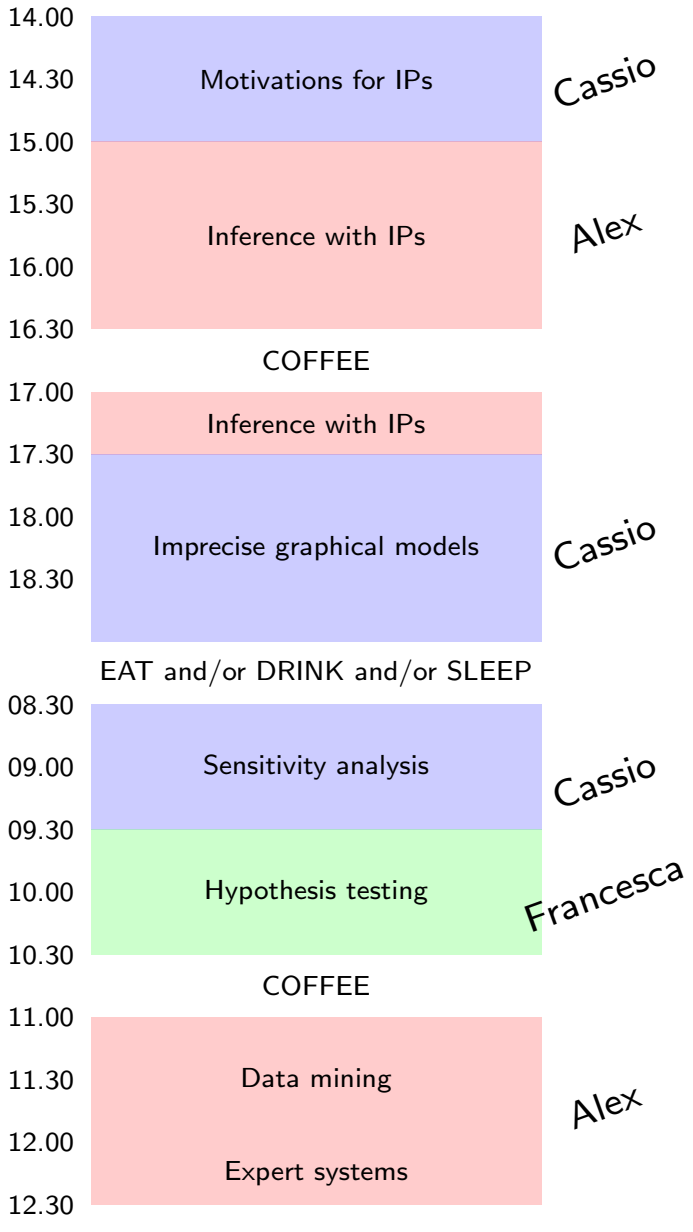
# Scheduling these two days

First Day

- Why IPs? Five motivating problems
- IPs in practice? Inference algorithms
- Many variables? IP graphical models

Second Day

- Sensitivity analysis
- Hypothesis testing
- Data mining
- Knowledge-based expert systems



# Motivating Imprecise Probability

- Proper treatment of missing data
- Reliable classification
- Sensitivity analysis
- Feature selection – "robust" statistical tests
- Representation of qualitative assessments
- Elicitation of expert knowledge

# R

- R: Programming language and environment tuned for statistical analysis.
- *http://www.r-project.org*
- Download and install it!

- `install.package("e1071")`
- `install.package("bnlearn")`
- `install.package("IDPSurvival")`
- `install.package("lpSolve")`
- `install.package("rcdd")`

# Missing Data – an example

- Patient presents symptoms that could be related to lung cancer.
- Physician can run tests for *Bronchitis* and do *X-rays*, as well as check for *Dyspnea*. However, (supposedly) they can only assess whether the patient is a *Smoker* by asking the patient themselves.
- Patient did not answer whether they are a smoker in the questionnare.
- (Hidden information: patient has a discount in their insurance because they declared *not* to be a smoker to the insurance company.)

Should *smoking* be ignored? Should it be marginalized out?
Should it be treated with (greater) care?

# Missing Data – another example

- Suppose we are given the following questionnaire.
- The possible answers are *bad*, *so-so*, *good*. It is also possible to leave it empty.

| What about? | Vlad | Barack | Roger | Maria | Penelope |
|---|---|---|---|---|---|

# Missing Data – another example

- Suppose we are given the following questionnaire.
- The possible answers are *bad*, *so-so*, *good*. It is also possible to leave it empty.

| What about? | Vlad | Barack | Roger | Maria | Penelope |
|---|---|---|---|---|---|
| Mr E | | | | | |
| Mr A | | | | | |
| Mr DC | | | | | |

Let's fill it in!

# Missing Data – another example

- Earlier today I gave the following questionnaire to three people, whose identity shall be kept anonymous.
- The only answer options are *bad* or *so-so*. It is also possible to leave it empty.

| What about? | Vlad | Barack | Roger | Maria | Penelope |
|---|---|---|---|---|---|
| Mr E | bad | so-so | | | bad |
| Mr A | bad | | so-so | good | so-so |
| Mr DC | bad | bad | so-so | good | |

# Missing Data – another example

- Earlier today I gave the following questionnaire to three people, whose identity shall be kept anonymous.
- The only answer options are *bad* or *so-so*. It is also possible to leave it empty.

| | Vlad | Barack | Roger | Maria | Penelope |
|---|---|---|---|---|---|
| Mr E | bad | so-so | GREAT | AMAZING | bad |
| Mr A | bad | GOOD | so-so | WHO IS? | so-so |
| Mr DC | bad | bad | so-so | good | GREAT |

"The only way to obtain unbiased estimation is to model missingness."

# Missing Data

- Should we consider missing as another category?
- Should we consider all possible completions of the data?
- Should we treat some missing values in a way, some in another?

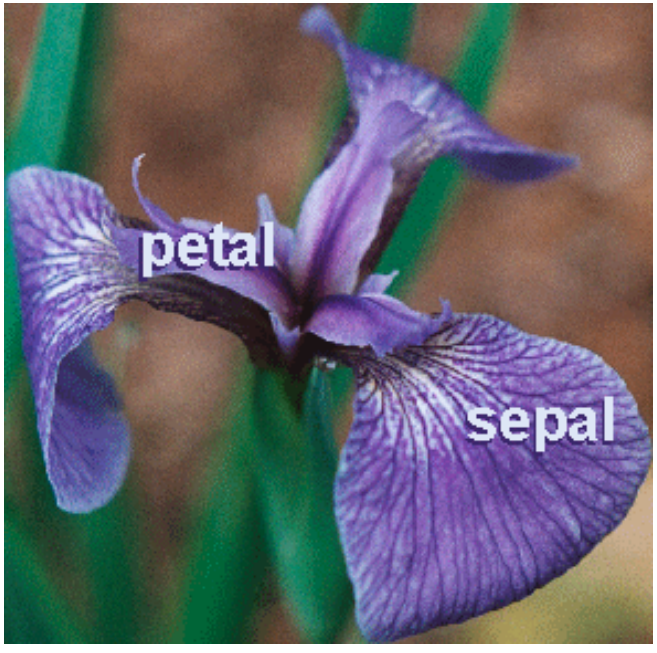
# Reliable classification

- Consider the following problem:
  - Objects contain some defining features (say $m$ of them) that (possibly) can be used to identify them.
  - Objects can be categorized into classes. The class of an object might be unknown to us.
- Given a collection of objects of known classes, build a model that can "guess" the class of an object of unknown class.
- Let us assume a log-linear model ($C$ class var, $F_i$ features):

$$P(C|F_1,\ldots,F_m) \propto P(C)\cdot\prod_{i=1}^{m} P(F_m|C)$$

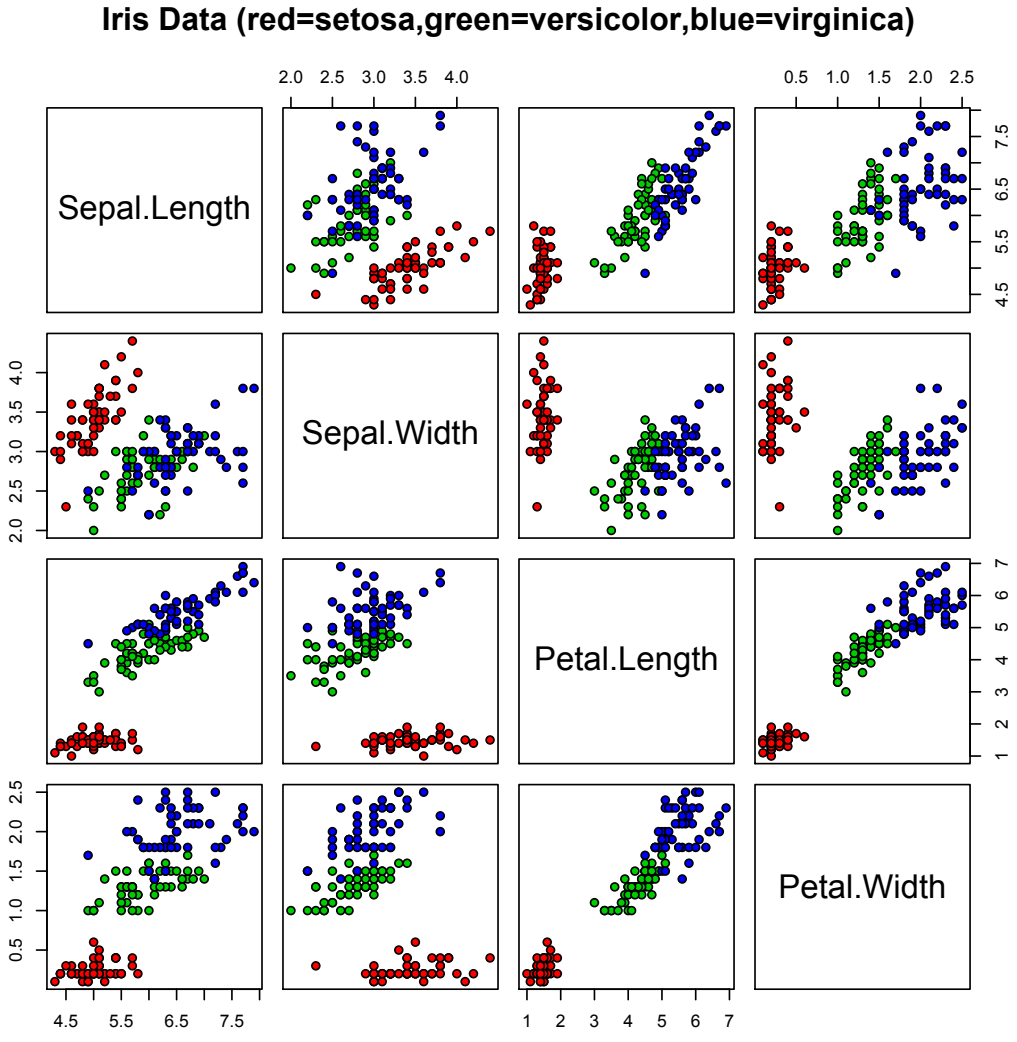
- Given $F_1,\ldots,F_m$, guessing the class can be done by taking $\max_C P(C|F_1,\ldots,F_m)$.
- Values $P(C)$ and $P(F_i|C)$ can be inferred using the collection of objects of known classes.

# Reliable classification - Iris example




http://mirlab.org/jang/books/dcpr/image/iris.gif

# Reliable classification - Iris example



# Iris example - classification1.txt

```
# http://www.bombonera.org/sipta/

options(width=300)
library(e1071)
data(iris)
#pairs(iris[1:4],
#  main="Iris Data (R=setosa,G=versicolor,B=virginica)",
#  pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
for(i in 1:ncol(iris)) if(is.numeric(iris[,i]))
  iris[,i] <- as.factor(iris[,i] > median(iris[,i]))
iris[1:10,]
iris.training <- iris[c(1:30,51:80,101:130),]
iris.testing <- iris[-c(1:30,51:80,101:130),]
```

# Iris example - classification2.txt

```
model <- naiveBayes(Species ~ ., laplace=1,
  data = iris.training)
prediction.classes <- predict(model, iris.testing)
prediction.probs <- predict(model, iris.testing, type='raw')
probs.max <- apply(prediction.probs, 1, max)

sum(prediction.classes == iris.testing$Species)/
   length(iris.testing$Species)
table(prediction.classes,iris.testing$Species)
summary(prediction.classes)
```

# Iris example - classification3.txt

```
cut <- 0.95
prediction.classes.high <- prediction.classes[probs.max > cut]
sum(prediction.classes.high ==
    iris.testing$Species[probs.max > cut])/
    length(prediction.classes.high)
table(prediction.classes.high,
         iris.testing$Species[probs.max > cut])

prediction.classes.low <- prediction.classes[probs.max <= cut]
sum(prediction.classes.low ==
     iris.testing$Species[probs.max <= cut])/
     length(prediction.classes.low)
table(prediction.classes.low,
     iris.testing$Species[probs.max <= cut])
```

# Reliable classification

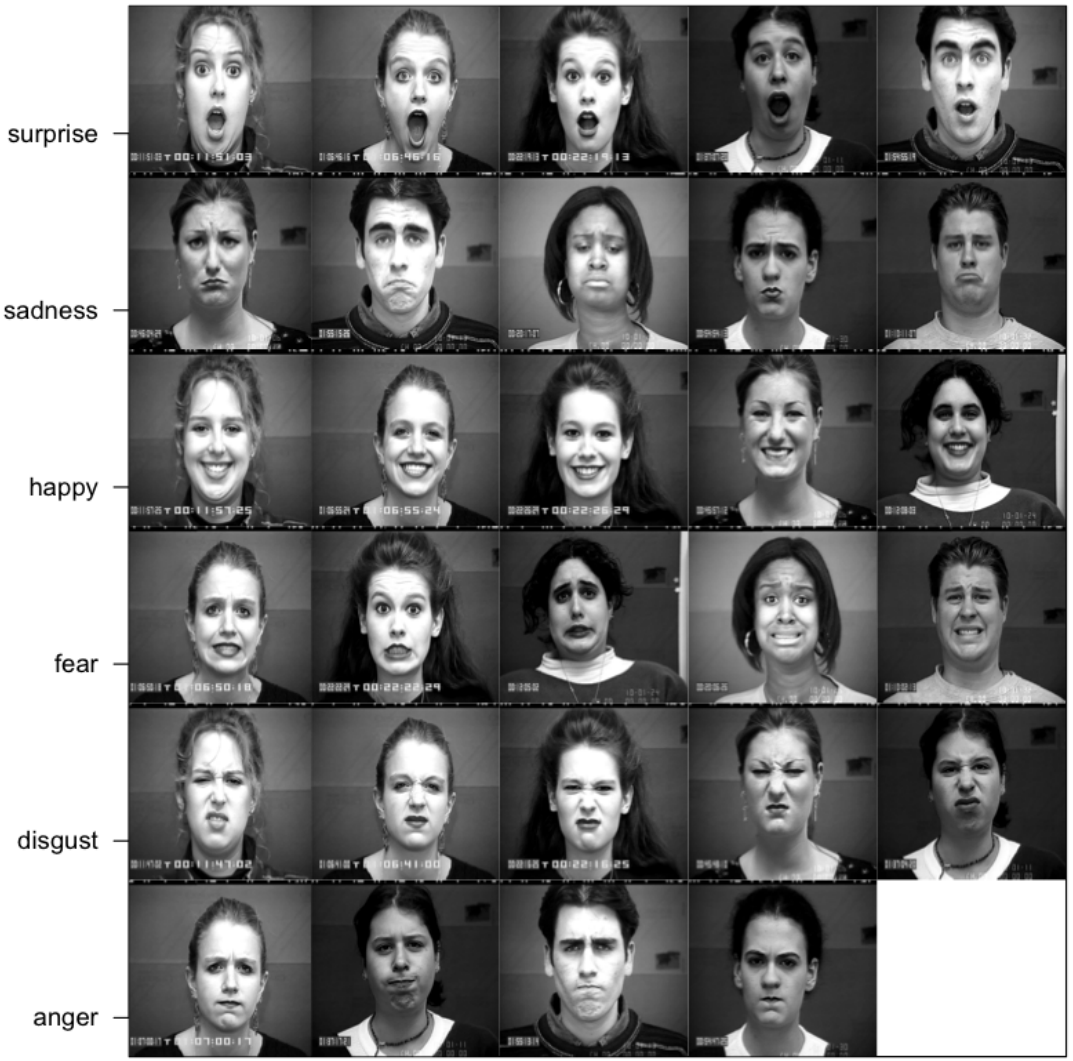
- Can we improve classification accuracy?
- Can we provide a subset of the classes that contains the correct one?
- Can we identify hard- and easy-to-classify instances?

- If probabilities are wrong, a simple cut-off or rejection rule might not be enough.

# Sensitivity Analysis

- Suppose that using a probabilistic model, we have reached a conclusion. Is this conclusion sensitive to modifications of the model?
- Usual procedure is to apply local modifications to the model and to check whether the conclusion remains inaltered.

# Sensitivity Analysis - an example

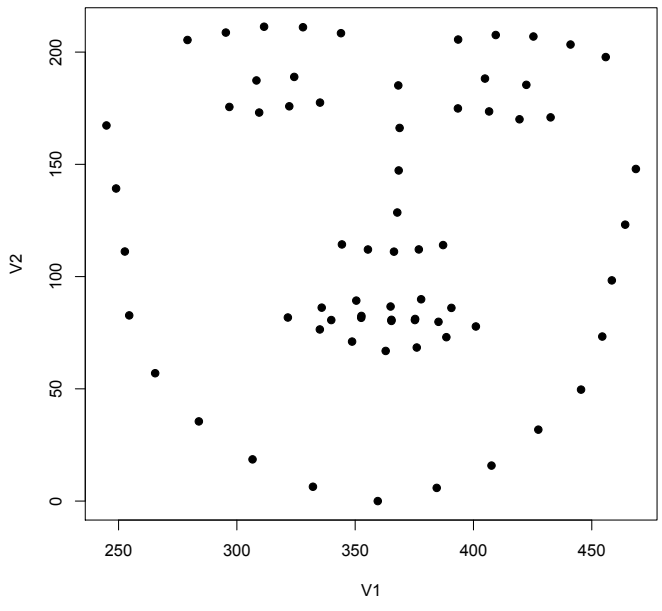



Cohn-Kanade (CK+) database, CVPR 2010.

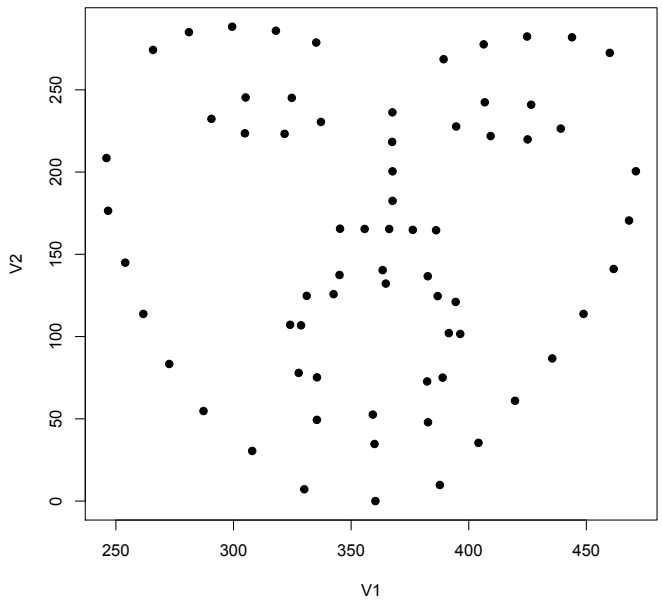
If you can, try to identify the best expression representing what Mr E thinks of Penelope.

# Sensitivity Analysis - an example

Anger



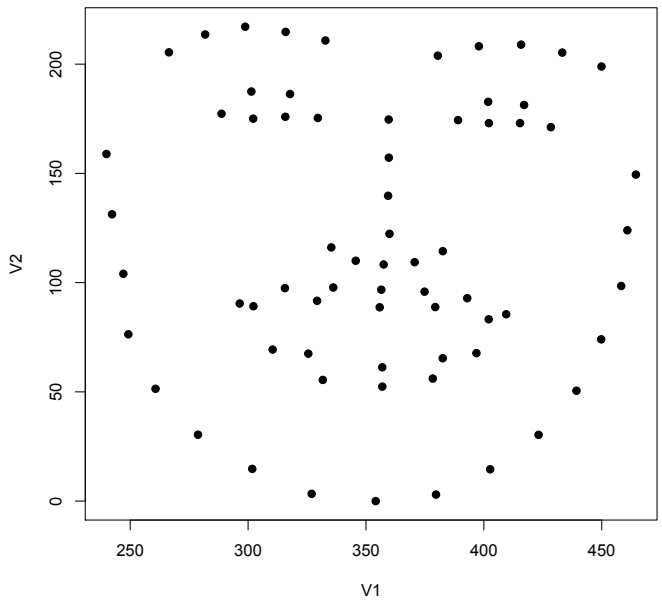

# Sensitivity Analysis - an example



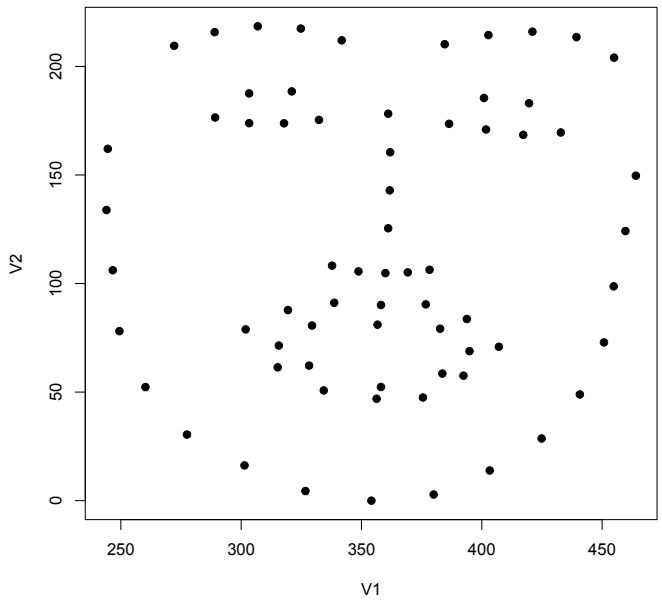

Surprise

# Sensitivity Analysis - an example

Happy




# Sensitivity Analysis - an example




Fear

# Sensitivity Analysis - an example

- We build a model with 23 Facial action units (facs) from the landmarks.
- We predict all these 23 facs.
- Standard techniques achieve about 90% accuracy.

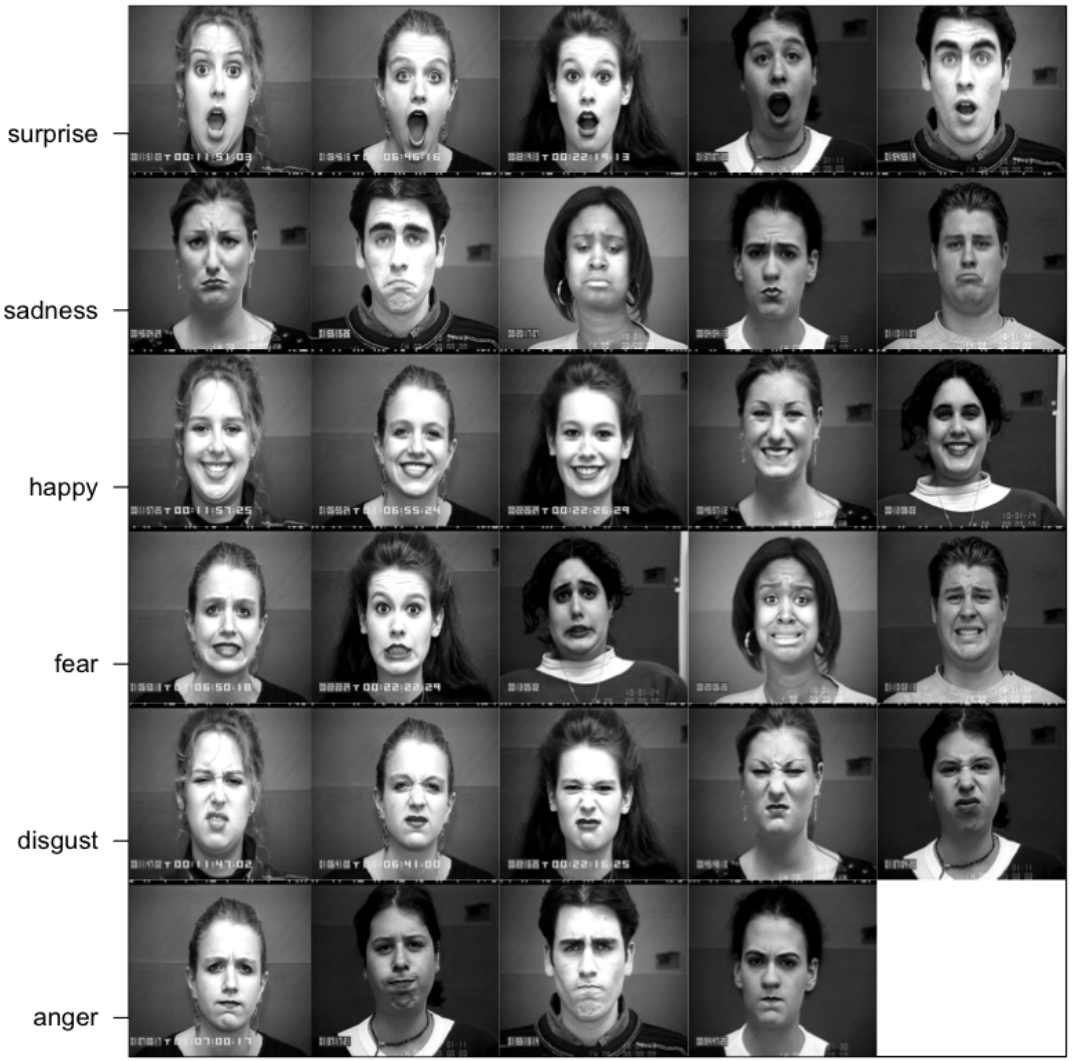
| AU Number | FACS Name | Muscular Basis |
|---|---|---|
| 0 | Neutral face | |
| 1 | Inner Brow Raiser | frontalis (pars medialis) |
| 2 | Outer Brow Raiser | frontalis (pars lateralis) |
| 4 | Brow Lowerer | depressor glabellae, depressor supercilii, corrugator supercilii |
| 5 | Upper Lid Raiser | levator palpebrae superioris, superior tarsal muscle |
| 6 | Cheek Raiser | orbicularis oculi (pars orbitalis) |
| 7 | Lid Tightener | orbicularis oculi (pars palpebralis) |
| 8 | Lips Toward Each Other | orbicularis oris |
| 9 | Nose Wrinkler | levator labii superioris alaeque nasi |
| 10 | Upper Lip Raiser | levator labii superioris, caput infraorbitalis |

Source: wikipedia.

# Sensitivity Analysis

- The model for expression recognition is quite complicate. Are the results reliable?
- If we employ a small modification in one parameter, would results change?
- If we allow all model parameters to vary within a region (near the estimated values), would results change?
- Some facial expressions are arguably easier to spot. Can we automatically identify that?

# Sensitivity Analysis - an example




Cohn-Kanade (CK+) database, CVPR 2010.

If you can, try to identify the best expression representing what Mr E thinks of Penelope.

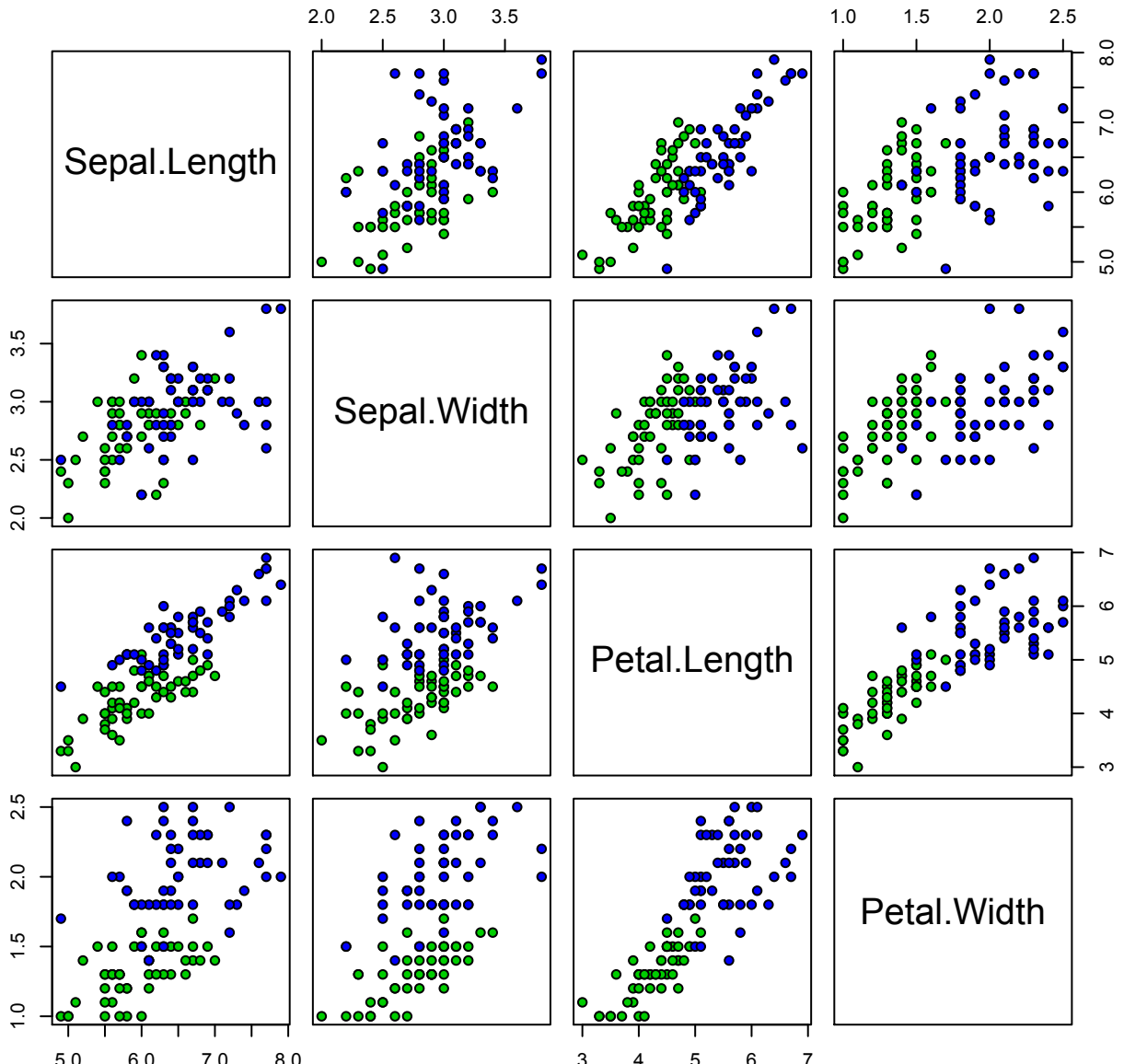
# "Robust" feature selection

- We are given a (potentially large) number of covariates and want to identify those which are useful to predict a binary response.
- For example, let us choose only some $F_i$ to include in the model:

$$P(C|F_1, \ldots, F_m) \propto P(C) \cdot \prod_{i=1}^{m} P(F_m|C)$$

- An usual procedure is to employ some statistical tests. An example is the Mann-Whitney u-test (a.k.a. Wilcoxon rank-sum test) to test whether $P(X > Y) \neq 0.5$ (or specific to one of the inequality sides), where $X$ and $Y$ represent some same characteristic in two populations.
- "MannWhitney is more robust than the Student's t-test", Wikipedia [citation needed]. "If it is written on wikipedia, then it is true", anonymous author.

# "Robust" feature selection - an example



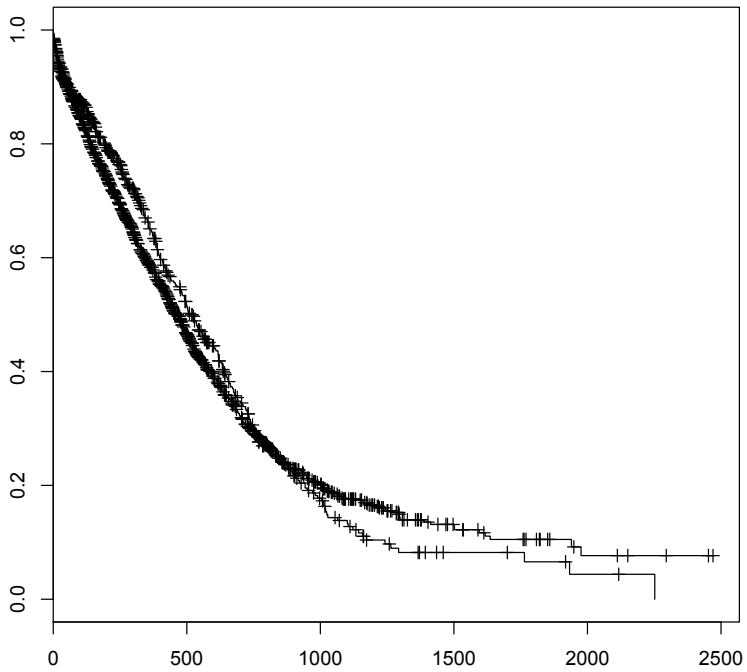
# "Robust" feature selection - an example - feature1.txt

```
data(iris)
#iris <- iris[iris$Species != 'setosa',]
#pairs(iris[1:4],
#  main="Iris Data (R=setosa,G=versicolor,B=virginica)",
#  pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
iris.versicolor <- iris[iris$Species == 'versicolor',]
iris.virginica <- iris[iris$Species == 'virginica',]
n = nrow(iris.virginica)

set.seed(1)
g1 <- sample.int(n,n/2)
g2 <- sample.int(n,n/2)

wilcox.test(iris.versicolor$Sepal.Width[g1],
   iris.virginica$Sepal.Width[g2],alternative='less')
wilcox.test(iris.versicolor$Sepal.Width[-g1],
   iris.virginica$Sepal.Width[-g2],alternative='less')
```

# "Robust" feature selection - an example

- Can we do feature selection in a principled and "robust" manner?

# "Robust" feature selection - yet another example




Survival of AIDS patients in Australian regions (Victoria and New South Wales).

# "Robust" feature selection - yet another example

- Log-rank (Mantel-Haenszel) test.
- Gehan-Wilcoxon (Peto & Peto modified) test.

# "Robust" feature selection - yet another example - feature2.txt

```
library(IDPSurvival)
data(Aids2)
Aids2[1:10,]
Aids2$death = Aids2$death - Aids2$diag
Aids2$status = 1*(Aids2$status=='D')
aids <- Aids2[(Aids2$state=='VIC')|(Aids2$state=='NSW') ,]
plot(survfit(Surv(aids$death,aids$status) ~ 1))
plot(survfit(Surv(aids$death,aids$status) ~ (aids$state=='VIC')))

testLR <- survdiff(Surv(aids$death,aids$status) ~
    (aids$state=='VIC'), rho = 0)
zLR <- sign(testLR$obs[1]-testLR$exp[1])*sqrt(testLR$chisq)
print(1-pnorm(zLR))
testPP <- survdiff(Surv(aids$death,aids$status) ~
    (aids$state=='VIC'), rho = 1)
zPP <- sign(testPP$obs[1]-testPP$exp[1])*sqrt(testPP$chisq)
print(1-pnorm(zPP))
```

# "Robust" statistical tests

- Can we tell whether the result of a statistical test is robust or not?
- Are usual tests calibrated?
- Can we come up with a measure of "robustness" for the test result?

# Eliciting expert knowledge / Qualitative assessments

- A tennis match between Maria and Penelope.
- Result of Maria after two sets: win, draw or loss?

## DETERMINISM

*Penelope does not know what is a racket, and Maria is a professional*

Maria (certainly) wins

$$\begin{matrix} P(Win) \\ P(Draw) \\ P(Loss) \end{matrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

## UNCERTAINTY

*Maria can only use the handle of the racket to hit. Win is two times more probable than draw, and this being three times more probable than loss*

$$\begin{matrix} P(Win) \\ P(Draw) \\ P(Loss) \end{matrix} = \begin{bmatrix} .6 \\ .3 \\ .1 \end{bmatrix}$$

## IMPRECISION

*Maria is blindfolded. Win is more probable than draw, and this is more probable than loss*

$$P(Win) \geq P(Draw)$$
$$P(Draw) \geq P(Loss)$$

$$\begin{matrix} P(Win) \\ P(Draw) \\ P(Loss) \end{matrix} = \begin{bmatrix} \alpha+\beta+\gamma \\ \alpha+\beta \\ \alpha \end{bmatrix}$$

$\forall \alpha, \beta, \gamma$ such that $\alpha \geq 0, \beta \geq 0, \gamma \geq 0,$

# From prob distributions to (linear) previsions (and back)

- In vitro fertilization: 3 embryos transfered, but only $X$ implanted

$$\Omega_X := \{0, 1, 2, 3\}$$

- Bayesian net to assess $P(X|age, \ldots)$ [Corani et al., 2012]

$$P(X = 0) = .75 \quad P(X = 1) = .21 \quad P(X = 2) = .035 \quad P(X = 3) = .005$$

- Complete model to (precisely) compute any expectation:

$$E_P[X] = .295 = .21 + .7 + .015 \text{ (mean)}$$

- Given a generic $f : \Omega_X \to \mathbb{R}$, a (linear) functional/prevision

$$E_P[f] := \sum_{x \in \Omega_X} P(x) \cdot f(x)$$

- Vice versa, given $E[f]$, compute $P(x) = E[(X = x)]$
- Representation theorem: $E[f] = E_P[f]$ [Seb] [de Finetti, 1974]

# From sets of distributions to (lower) previsions (and back)

- A (imprecise/credal) network used to assess bounds on $P(X)$:

$$P(X=0) \geq .7 \quad P(X=1) \leq .25 \quad P(X=2) \leq .05 \quad P(X=3) \leq .01$$

- Only the bounds of the expectations can be computed

$$\underline{E}[f] = \min_{\substack{\sum_{x\in\Omega} P(x)=1 \\ P(x)\geq 0 \forall x\in\Omega_X \\ P(X=0)\geq .7\ P(X=1)\leq .25 \\ P(X=2)\leq .05\ P(X=3)\leq .01}} \sum_{x\in\Omega_X} P(x)\cdot f(x) \qquad \begin{aligned} \underline{E}(X) &= .00 \\ \overline{E}(X) &= .37 \end{aligned}$$

- Not anymore linear functional. E.g., prob of no triplet?

$$\underline{E}[X<3] = .99 > .70 = \underline{E}[(X=0)] + \underline{E}[(X=1)] + \underline{E}[(X=2)]$$

- Vice versa? Given $\underline{E}[f]$, compute $\underline{P}(x) := \underline{E}[(X=x)]$ and $\overline{P}(x)$

  The functional induced by these constraints is $\underline{E}'[f] \leq \underline{E}[f]$

# Credal sets [Levi, 1980]

- Imprecise knowledge about $X$ as a set $K(X)$ of distributions
- $K(X)$ induces the functional (coherent lower prev) [Walley, 1991]

$$\underline{E}_K[f] := \min_{P(X)\in K(X)} \sum_{x\in\Omega_X} P(x)\cdot f(x)$$

- $K(X)$ assumed to be convex (and closed)! Why?

  Convexification does not affect the optimum of a linear function
- $K(X)$ is a very general uncertainty model called credal set [Didier]
- Assume $K(X)$ induced by a finite number of linear constraints

  The set of extreme points $\mathrm{ext}[K(X)]$ has finite cardinality too
- The task is a LP and the solution is on an extreme point!

$$\underline{E}_K[f] = \min_{P(X)\in \mathrm{ext}[K(X)]} \sum_{x\in\Omega_X} P(x)\cdot f(x)$$

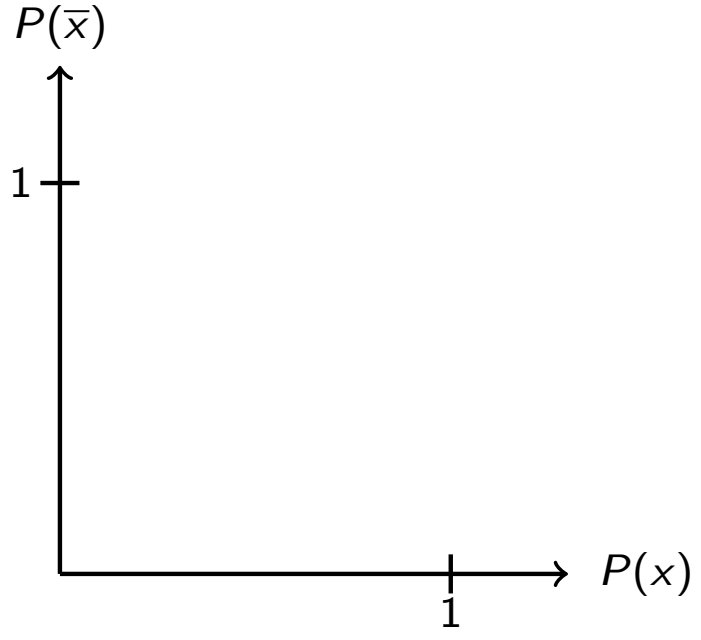
# Exercise #1 (`ex1.r`)

*Consider the quaternary variable $X$ in the fertilization example and the assessment $P(X)$ (precise) and $K(X)$ (imprecise)*

1. Compute $\mu_X := E_P(X)$ (mean) and $\sqrt{E_P[(\mu_X - X)^2]}$ (std dev)
2. Compute $\underline{E}_K(X)$ and $\overline{E}_K(X)$
3. Check sublinearity of $\underline{E}_K(X)$ by considering the prob of no triplets
4. Compute $\mathrm{ext}[K(X)]$
5. Repeat the (imprecise) computations using the extreme points

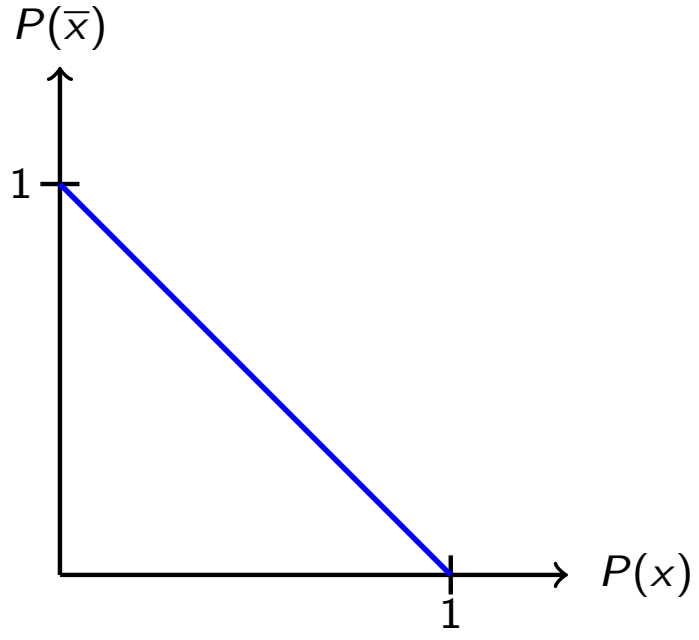
# Credal sets (CSs) over Boolean variables

- Boolean $X$, values in $\Omega_X = \{x, \overline{x}\}$
- Precise $P(X) = \begin{bmatrix} p \\ 1-p \end{bmatrix}$ $p \in [0,1]$



# Credal sets (CSs) over Boolean variables

- Boolean $X$, values in $\Omega_X = \{x, \overline{x}\}$
- Precise $P(X) = \begin{bmatrix} p \\ 1-p \end{bmatrix}$ $p \in [0,1]$
- Biggest CS? Whole simplex $K_v(X)$
- $K_v(X)$ has 2 extreme points

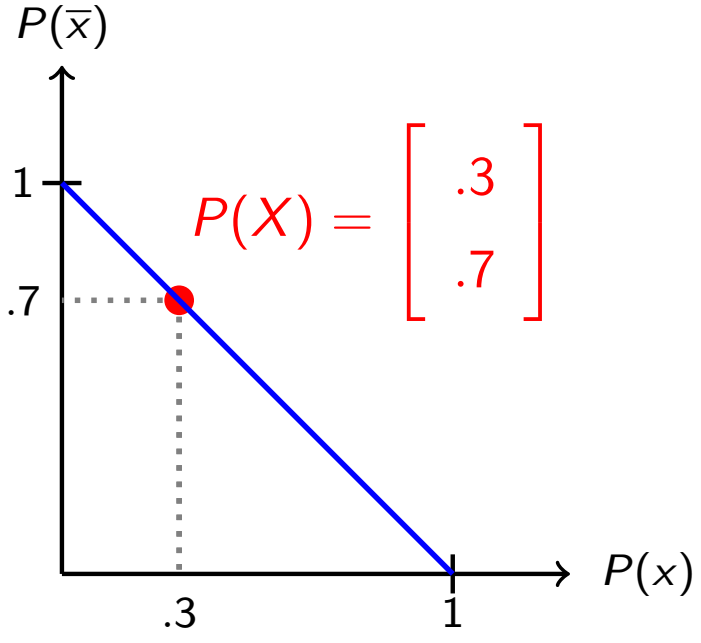


# Credal sets (CSs) over Boolean variables

- Boolean $X$, values in $\Omega_X = \{x, \overline{x}\}$
- Precise $P(X) = \begin{bmatrix} p \\ 1-p \end{bmatrix}$ $p \in [0,1]$
- Biggest CS? Whole simplex $K_v(X)$
- $K_v(X)$ has 2 extreme points
- True for any CS (convexity in 1D)
- Combinatorial approach appealing!

  Intervals are fully general
- Binarization techniques to describe a generic model with Boolean variables

  [Antonucci et al., 2008]




$\underline{E}_K[(2,-1)] = 0.2$

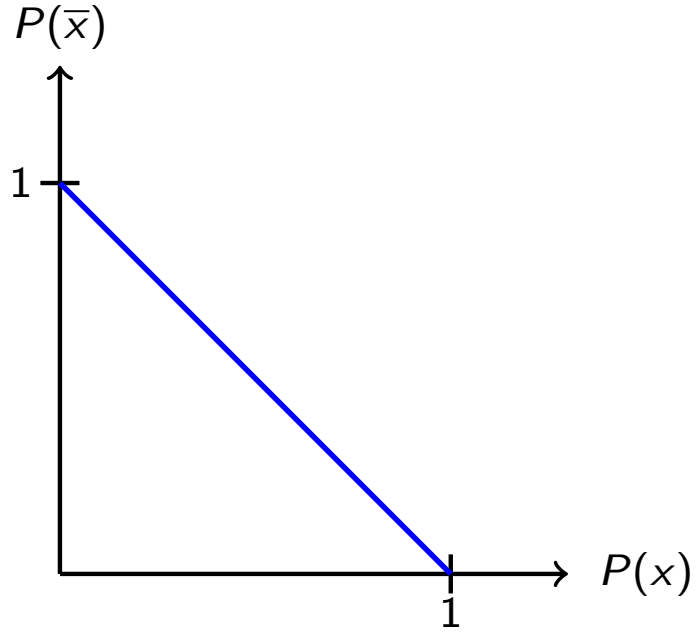
$\overline{E}_K[(2,-1)] = 1.1$

# Credal sets (CSs) over Boolean variables

- Boolean $X$, values in $\Omega_X = \{x, \overline{x}\}$
- Precise $P(X) = \begin{bmatrix} p \\ 1-p \end{bmatrix}$ $p \in [0,1]$
- Biggest CS? Whole simplex $K_v(X)$
- $K_v(X)$ has 2 extreme points
- True for any CS (convexity in 1D)
- Combinatorial approach appealing!

  Intervals are fully general
- Binarization techniques to describe a generic model with Boolean variables

  [Antonucci et al., 2008]




$\underline{E}_K[(2,-1)] = 0.2$

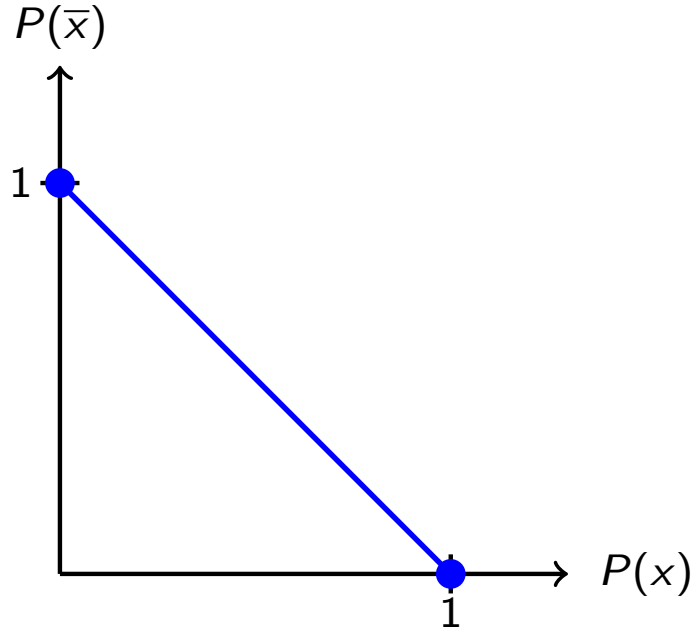
$\overline{E}_K[(2,-1)] = 1.1$

# Credal sets (CSs) over Boolean variables

- Boolean $X$, values in $\Omega_X = \{x, \overline{x}\}$
- Precise $P(X) = \begin{bmatrix} p \\ 1-p \end{bmatrix}$ $p \in [0,1]$
- Biggest CS? Whole simplex $K_v(X)$
- $K_v(X)$ has 2 extreme points
- True for any CS (convexity in 1D)
- Combinatorial approach appealing!

  Intervals are fully general
- Binarization techniques to describe a generic model with Boolean variables

  [Antonucci et al., 2008]



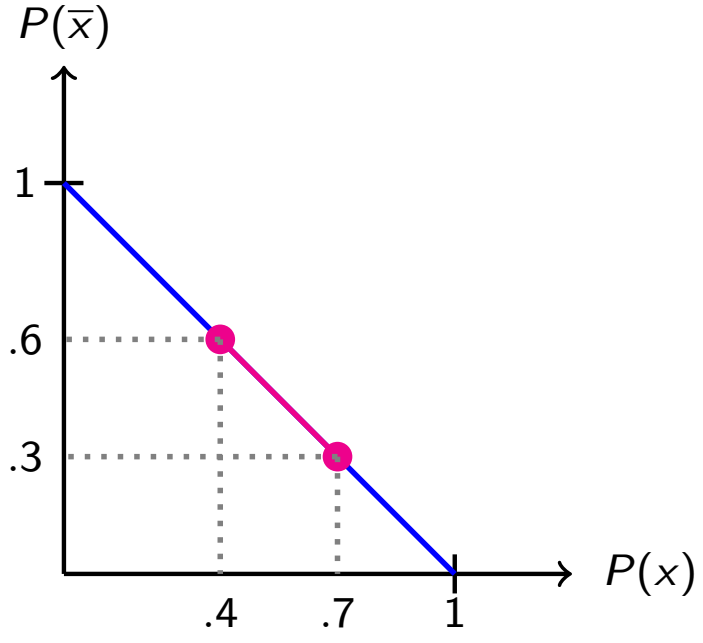

$\underline{E}_K[(2,-1)] = 0.2$

$\overline{E}_K[(2,-1)] = 1.1$

# Credal sets (CSs) over Boolean variables

- Boolean $X$, values in $\Omega_X = \{x, \overline{x}\}$
- Precise $P(X) = \begin{bmatrix} p \\ 1-p \end{bmatrix}$ $p \in [0,1]$
- Biggest CS? Whole simplex $K_v(X)$
- $K_v(X)$ has 2 extreme points
- True for any CS (convexity in 1D)
- Combinatorial approach appealing!
  Intervals are fully general
- Binarization techniques to describe a generic model with Boolean variables

  [Antonucci et al., 2008]




$$K(X) \equiv \{P(X) | .4 \leq P(x) \leq .7\}$$

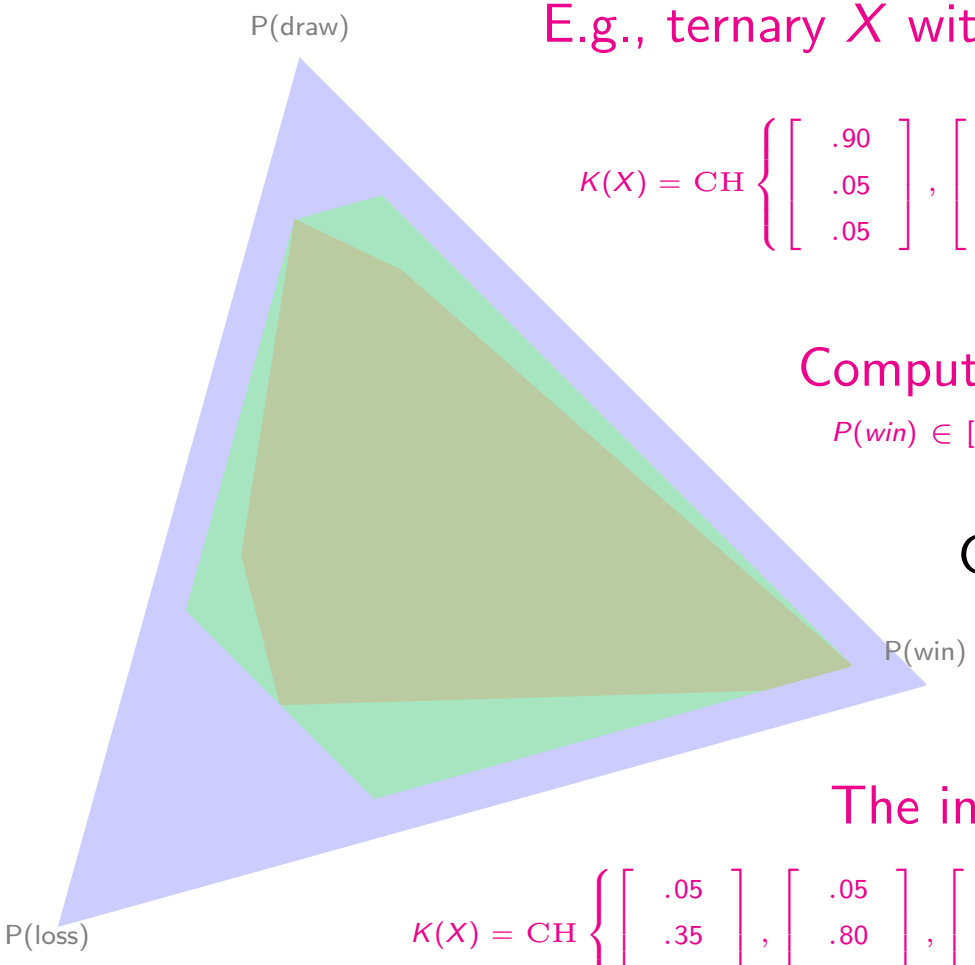
$$\underline{E}_K[(2,-1)] = 0.2$$
$$\overline{E}_K[(2,-1)] = 1.1$$

# But if $|\Omega_X| > 2 \ldots$

No bounds on the number of extreme points!

E.g., ternary $X$ with $\Omega_X = \{win, draw, loss\}$

$$K(X) = \mathrm{CH}\left\{\begin{bmatrix} .90 \\ .05 \\ .05 \end{bmatrix}, \begin{bmatrix} .80 \\ .05 \\ .15 \end{bmatrix}, \begin{bmatrix} .20 \\ .20 \\ .60 \end{bmatrix}, \begin{bmatrix} .10 \\ .40 \\ .50 \end{bmatrix}, \begin{bmatrix} .05 \\ .80 \\ .15 \end{bmatrix}, \begin{bmatrix} .20 \\ .70 \\ .10 \end{bmatrix}\right\}$$




Compute bounds on probabilities

$P(win) \in [.05, .90]$ $P(draw) \in [.05, .80]$ $P(loss) \in [.05, .60]$

CSs from intervals not fully general

$\underline{P}(X)/\overline{P}(X)$ note expressive enough

[Quique]

The induced CS is larger

$$K(X) = \mathrm{CH}\left\{\begin{bmatrix} .05 \\ .35 \\ .60 \end{bmatrix}, \begin{bmatrix} .05 \\ .80 \\ .15 \end{bmatrix}, \begin{bmatrix} .15 \\ .80 \\ .05 \end{bmatrix}, \begin{bmatrix} .35 \\ .05 \\ .60 \end{bmatrix}, \begin{bmatrix} .90 \\ .05 \\ .05 \end{bmatrix}\right\}$$

# Exercise #2 (`ex2.r`)

1. Brazil winning the Wolrd Cup?

   As an imprecise Brazilian, Cassio's probs are $.4 \leq P(win) \leq .7$

   The gamble is: +2EUR if Brazil wins and pay -1EUR if not

   Cassio's lowest (selling) and upper (buying) price?

2. Can you find a gamble whose lower expectation is different

   for the two credal sets considered in the previous slide?

   And an indicator?

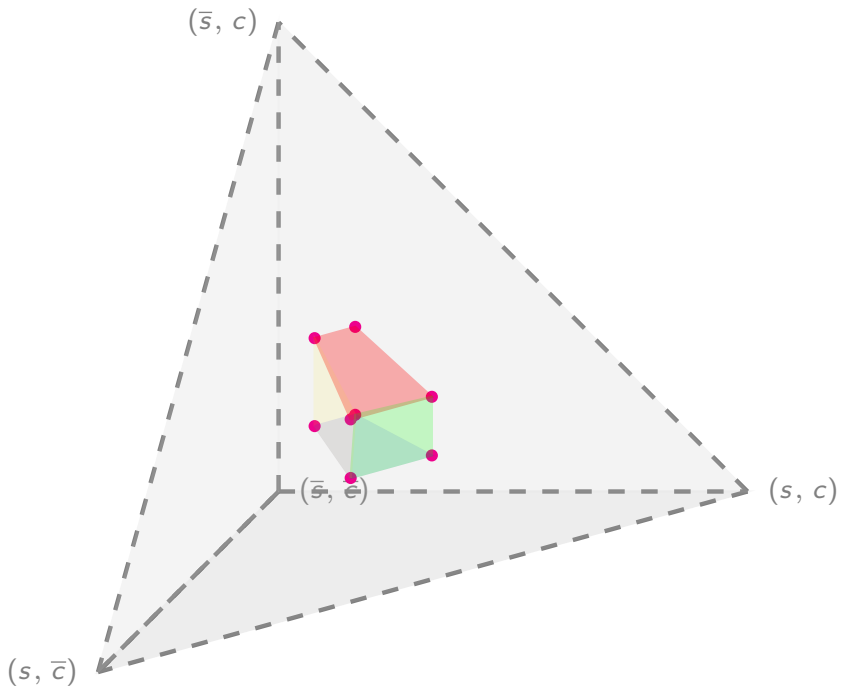
# Multivariate models

- 8 "Bayesian" physicians, each assessing $P_j(S, C)$

| $j$ | $P_j(s, c)$ | $P_j(s, \overline{c})$ | $P_j(\overline{s}, c)$ | $P_j(\overline{s}, \overline{c})$ |
|---|---|---|---|---|
| 1 | 2/16 | 2/16 | 3/16 | 9/16 |
| 2 | 2/16 | 2/16 | 6/16 | 6/16 |
| 3 | 3/16 | 1/16 | 3/16 | 9/16 |
| 4 | 3/16 | 1/16 | 6/16 | 6/16 |
| 5 | 4/16 | 4/16 | 2/16 | 6/16 |
| 6 | 4/16 | 4/16 | 4/16 | 4/16 |
| 7 | 6/16 | 2/16 | 2/16 | 6/16 |
| 8 | 6/16 | 2/16 | 4/16 | 4/16 |

- No second order information? Be cautious, take convex hull $K(S, C) = \mathrm{CH}\left\{P_j(S, C)\right\}_{j=1}^{8}$

Two Boolean variables **S**moker, Lung **C**ancer




3D projection of a 4D object

# Marginalization in multivariate CSs

- Prob of cancer based on $P(S, C)$? This is a marginalization

$$P(c) = P(s, c) + P(\overline{s}, c) \quad [\text{e.g., } P_1(c) = \tfrac{5}{16}]$$

- Based on $K(S, C)$? Just the same (but elementwise)

$$K(C) = \left\{ P'(C) \;\middle|\; \begin{array}{l} P'(c) = P(s, c) + P(\overline{s}, c) \quad \forall c \\ \forall P(S, C) \in K(S, C) \end{array} \right\}$$

- In practice? Marginalize only extremes and take the convex hull!

$$K(C) = \mathrm{CH} \left\{ P'(C) \;\middle|\; \begin{array}{l} P'(c) = P(s, c) + P(\overline{s}, c) \quad \forall c \\ \forall P(S, C) \in \mathrm{ext}[K(S, C)] \end{array} \right\}$$

- This is used to implement $\underline{E}_{K(C)}$ [e.g., $\underline{P}(c) = \frac{5}{16}$ $\overline{P}(c) = \frac{10}{16}$]

# Conditioning in multivariate CSs

- Given $P(S, C)$, prob of cancer for smokers? Conditioning!

  If $P(s) > 0$, Bayes' rule says $P(c|s) = \frac{P(s,c)}{P(s)}$ [e.g., $P_1(c|s) = .5$]

- Given $K(S, C)$? Generalized Bayes' rule! [Quique] [Walley, 1991]

  Just the same (but elementwise)

$$K(C|s) = \left\{ P'(C|s) \;\middle|\; \begin{array}{l} P'(c|s) = \frac{P(s,c)}{P(s)} \quad \forall c \\ \forall P(S, C) \in K(S, C) \end{array} \right\}$$

- In practice? Condition the extremes and take the convex hull

$$K(C|s) = \mathrm{CH} \left\{ P'(C|s) \;\middle|\; \begin{array}{l} P'(c|s) = \frac{P(s,c)}{P(s)} \quad \forall c \\ \forall P(S, C) \in \mathrm{ext}[K(S, C)] \end{array} \right\}$$ [e.g., $\overline{P}(c|s) = \frac{3}{4}$]

- BR requires $P(s) > 0$, for GBR this corresponds to $\underline{P}(s) > 0$

  $\overline{P}(s) > 0$ is ok too: ignore a $P$ if $P(s) = 0$ (regular extension)

- $K(C|S)$ is a collection of CSs (one $\forall$ possible conditioning event)

# Back to the joint (marginal extension)

- Precise case: $P(s,c) = P(c|s) \cdot P(s)$ for each joint state

$$P(S,C) := P(C|S) \otimes P(S)$$

- Imprecise case? Given $K(C|S)$ and $K(S)$ build a joint CS

  Elementwise combination

$$K(C|S) \otimes K(X) := \left\{ P'(S,C) \;\middle|\; \begin{array}{ll} P'(s,c) = P(c|s) \cdot P(s) & \\ \forall (s,c) \quad \forall P(S,C) \in K(S,C) & \end{array} \right\}$$

- $\otimes$ in practice? Combining (all combinations) of the extremes + CH
- A round trip: start from $K(S,C)$, compute $K(C|S)$ and $K(S)$, then $K'(C,S) = K(C|S) \otimes K(S)$

  If $K(C,S) \equiv K'(C,S)$, $K(C)$ and $K(C|S)$ jointly coherent

  In general only $K(C,S) \subseteq K'(C,S)$ (in our example coincide)

- $\otimes$ + marg $\Rightarrow$ (set-)valuation algebra [Kohlas][Mauá et al., 2012]

# Exercise #3 (`ex3.r`)

- Consider the CS induced by the (convexification of the) 8 physicians
  1. Compute $K(C)$, $K(S)$, and $K(C|S)$,
  2. Check $K(S, C) = K(S) \otimes K(C|S)$

# (The) Independence (Zoo)

- A guy (working for Marlboro) says: *"C and S are independent"*
- He's a precise guy [he has his own $P(C,S)$] What does it means?

Stochastic independence
$P(c) \cdot P(s) = P(c,s)\ \forall c,s$ $\Leftrightarrow$ Stochastic irrelevance
$P(C|s) = P(c)\ \forall s$

- But the Marlboro guy is imprecise! [he has $K(C,S)$]

Strong independence
$P(c) \cdot P(s) = P(c,s)\ \forall c,s$
$\forall P(C,S) \in \mathrm{ext}[K(C,S)]$ (strong)
$\forall P(C,S) \in K(C,S) \Rightarrow$ non-convex
$\Rightarrow$ Epistemic irrelevance
$K(C|s) = K(c)\ \forall s$

Strong independence is symmetric, epistemic irrelevance is not

*Every notion of independence/irrelevance has a conditional version*

# Extensions based on independence

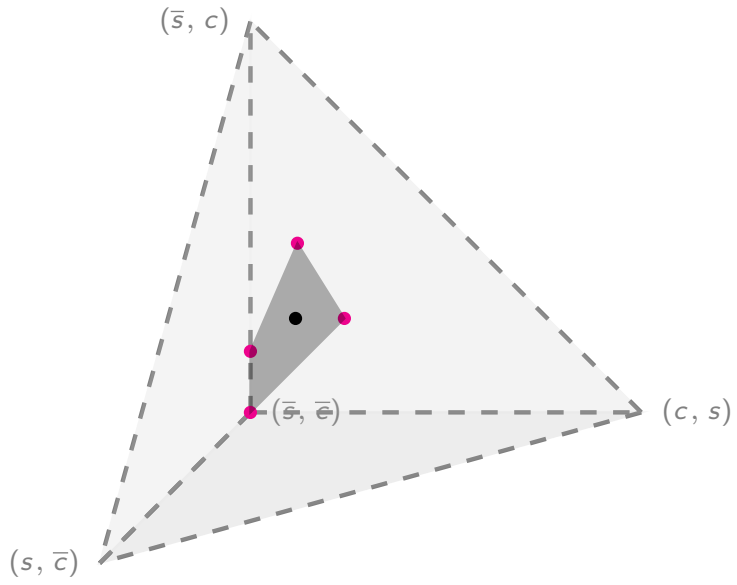
- Precise marginals $P(C)$ and $P(S)$ $[P(c) = \frac{1}{2}$ and $P(s) = \frac{3}{8}]$
- Enough to build a consistent $P(S, C)$ s.t. $C$ and $S$ (stoch) indep
- Imprecise marginals $K(C)$ and $K(S)$ $[P(c) \in [\frac{5}{16}, \frac{5}{8}]$, $P(s) \in [\frac{1}{4}, \frac{1}{2}]]$
- Build the smallest joint CS consistent with $K(C)$ and $K(S)$ s.t.
  - $C$ and $S$ are strongly independendent (strong extension)
  - $S$ is epistemically irrelevant to $C$ (epistemic extension)

# Building the strong extension

- Extreme points of the extension are combinations of the extreme points of the marginals [Antonucci, 2008]

$$\text{ext}[K(C)] = \text{CH}\left\{\begin{bmatrix} \frac{5}{16} \\ \frac{13}{16} \end{bmatrix}, \begin{bmatrix} \frac{5}{8} \\ \frac{3}{8} \end{bmatrix}\right\}$$

$$\text{ext}[K(S)] = \text{CH}\left\{\begin{bmatrix} \frac{1}{4} \\ \frac{3}{4} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}\right\}$$



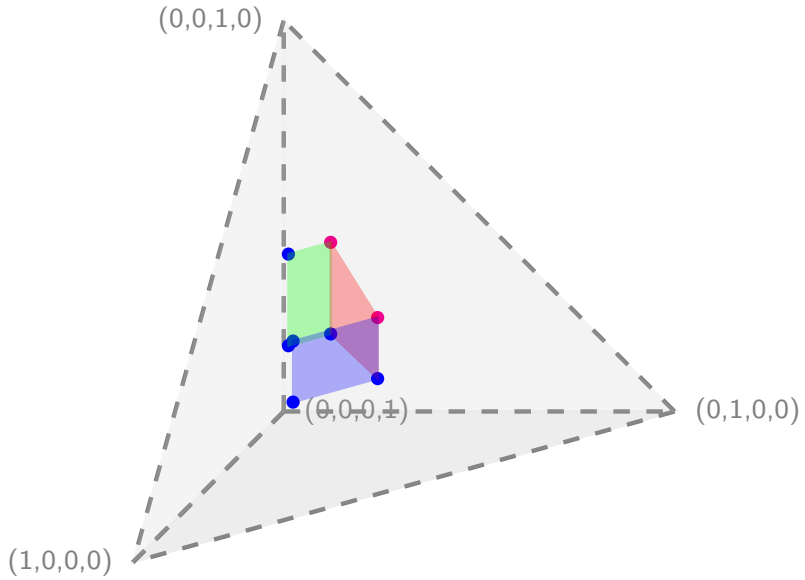

$$K_{SE}(C,S) = \text{CH}\left\{\begin{pmatrix} \frac{5}{64} \\ \frac{13}{64} \\ \frac{15}{64} \\ \frac{39}{64} \end{pmatrix}, \begin{pmatrix} \frac{10}{64} \\ \frac{6}{64} \\ \frac{30}{64} \\ \frac{18}{64} \end{pmatrix}, \begin{pmatrix} \frac{10}{64} \\ \frac{26}{64} \\ \frac{10}{64} \\ \frac{26}{64} \end{pmatrix}, \begin{pmatrix} \frac{20}{64} \\ \frac{12}{64} \\ \frac{20}{64} \\ \frac{12}{64} \end{pmatrix}\right\}$$

# Building the epistemic extension

- Irrelevance by linear constraints

  [Mauá et al., 2014]

  - $\frac{5}{16} = \underline{P}(c) \leq P(c|s) \leq \overline{P}(c) = \frac{5}{8}$
    $\Rightarrow \frac{5}{16} \cdot P(s) \leq P(s,c) \leq \frac{5}{8} \cdot P(s)$
    Where $P(s) = P(c,s) + P(\overline{c},s)$
  - Likewise for $P(c|\overline{s})$
  - Don't forget $\frac{5}{16} \leq P(c) \leq \frac{5}{8}$ and $\frac{1}{4} \leq P(s) \leq \frac{1}{2}$




- strong $\subseteq$ epistemic extension
- More extreme points: combinatorial approach less appealing
- Use previsions or desirable gambles

  [Gert & Jasper]

# Exercise #4 (`ex4.r`)

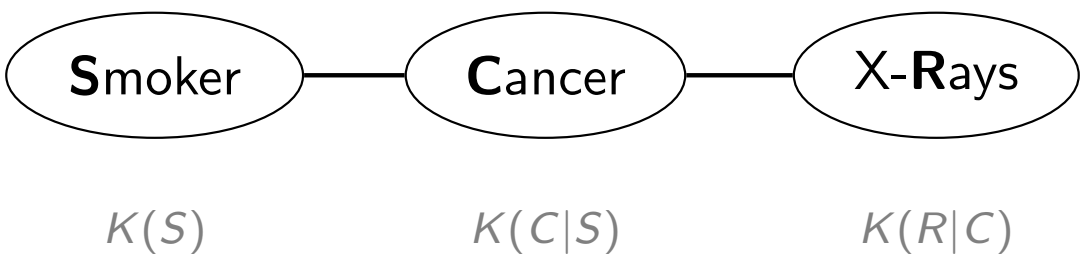
1. Compute $K_{SE}(S, C)$
2. Compute $K_{EE}(S, C)$

# Multivariate models with independence

- A third Boolean variable: X-**R**ays to reveal lung cancer
- For non-Marlboro people too given cancer no link smoke/X-rays
- IP language: given $C$, $S$ and $R$ strongly independent
- Marginal extension (iterated two times)

  $$K(S,C,R) = K(R|S,C) \otimes K(S,C) = K(R|S,C) \otimes K(C|S) \otimes K(S)$$
- Independence implies irrelevance: given $C$, $S$ irrelevant to $X$

  $$K(S,C,R) = K(R|C) \otimes K(C|S) \otimes K(S)$$
- Global model decomposed in 3 "local" models
- A graphical model! Independence make specification compact



# Strong extension with conditional independence

- 3 "local" $K(S)$, $K(C|S)$, $K(R|C)$ + indep statement
- $K(R|C)$ s.t. $P(r|c) \in [.9, .95]$ $P(\neg r|\neg c) \in [.9, .95]$
- $K_{SE}(S, C, R) := K(S) \otimes K(C|S) \otimes K(R|S)$
  the smallest consistent CS satisfying independence
- This generalizes the notion of strong extension
- $\mathrm{ext}[K_{SE}(S, C, R)]$? Combination of the local extreme points
- $2 \cdot 4 \cdot 4 = 32$ points (some of them in the convex hull)
- number of extreme points exponential wrt the number of variables

# Exact inference: brute-force combinatorial approach

- Demo inference in the trivariate model: $\underline{P}(r)$?
- Inference w.r.t. to the strong extension $K_{SE}(S, C, R)$

$$\underline{P}(r) = \min_{P(S,C,R) \in K_{SE}(S,C,R)} [P(s, c, r) + P(\overline{s}, c, r) + P(s, \overline{c}, r) + P(\overline{s}, \overline{c}, r)]$$

- Linear problem (in the joint space): solution on an extreme point

$$\underline{P}(r) = \min_{P(S,C,R) \in \mathrm{ext}[K_{SE}(S,C,R)]} [P(s, c, r) + P(\overline{s}, c, r) + P(s, \overline{c}, r) + P(\overline{s}, \overline{c}, r)]$$

$$\underline{P}(r) = \min_{i=1,\ldots,32} [P_i(s, c, r) + P_i(\overline{s}, c, r) + P_i(s, \overline{c}, r) + P_i(\overline{s}, \overline{c}, r)] = .288125$$

- Two levels of complexity:
  - Bayesian: exponential # number of sums
  - Credal: exponential # number of extreme points

# Exact inference: multilinear optimization

- Linearly constrained optimization of a multilinear function

$$\underline{P}(r) = \min_{\substack{\frac{1}{4} \le P(s) \le \frac{1}{2} \\ \frac{1}{2} \le P(c|s) \le \frac{3}{4}, \frac{1}{4} \le P(c|\bar{s}) \le \frac{1}{2} \\ \frac{1}{100} \le P(r|c) \le \frac{1}{10}, \frac{9}{10} \le P(r|\bar{c}) \le \frac{99}{100}}} \sum_{s,c} [P(s) \cdot P(c|s) \cdot P(r|c)]$$

- Multilinear solvers can be used,

  but the problem has degree equal to the number of variables

- Make it bilinear (i.e., degree=2) by (symbolic) variable elimination [de Campos & Cozman, 2008]

$$\underline{P}(r) = \min_{\substack{P(c)=\sum_s P(s)\cdot P(c|s) \\ P(\bar{c})=\sum_s P(s)\cdot P(\bar{c}|s) \\ \cdots}} \sum_c [P(r|c) \cdot P(c)]$$

- Write the problem in AMPL and solve it with CPLEX

- an "elimination" order over the variable is needed

# Linearizing the multilinear task

- Additional precise constraints for "local" probs (not for $R$)
- E.g., $P(s) = \frac{1}{4}$, $P(c|s) = \frac{3}{4}$ $P(c|\bar{s}) = \frac{1}{4}$

$$\underline{P}'(r) = \min_{\substack{P(s)=\frac{1}{4}, P(c|s)=\frac{3}{4}, P(c|\bar{s})=\frac{1}{4} \\ \frac{9}{10} \leq P(r|c) \leq \frac{99}{100}, \frac{9}{10} \leq P(\bar{r}|\bar{c}) \leq \frac{99}{100}}} \sum_{s,c} [P(s) \cdot P(c|s) \cdot P(r|c)] = .34375$$

- More constraints $\Rightarrow$ inner approx $\underline{P}'(r) \geq \underline{P}(r)$
- The solution gives me extreme points for $S$, give the freedom to $C$!
- Iterative approximate procedure [Antonucci et al., 2013]

# Interval propagation

- $\underline{P}(r) = \min_{P(c)} \sum_r \underline{P}(r|c) \cdot P(c)$
- $\underline{P}(c) = \min_{P(s)} \sum_s \underline{P}(c|s) \cdot P(s)$ and $\overline{P}(c) = \max_{P(s)} \sum_s \overline{P}(c|s) \cdot P(s)$
- Exact for Boolean variables [Zaffalon],

  (very) approximate (outer) in general [Tessem]

$$\underline{P}(c) = \min_{P(S) \in \mathrm{ext}[K(S)]} P(s) \cdot \underline{P}(c|s) + [1 - P(s)] \cdot \underline{P}(c|\overline{s}) =$$
$$= k + \min_{P(S) \in \mathrm{ext}[K(S)]} P(s)[\underline{P}(c|s) - \underline{P}(c|\overline{s})]$$

As $\frac{1}{4} = \underline{P}(c|\overline{s}) < \underline{P}(c|s) = \frac{1}{2}$,
the minimum is achieved when $P(s) = \frac{1}{4}$ (minimum value)

We have $\underline{P}(c) = 0.3125$ (exact!)

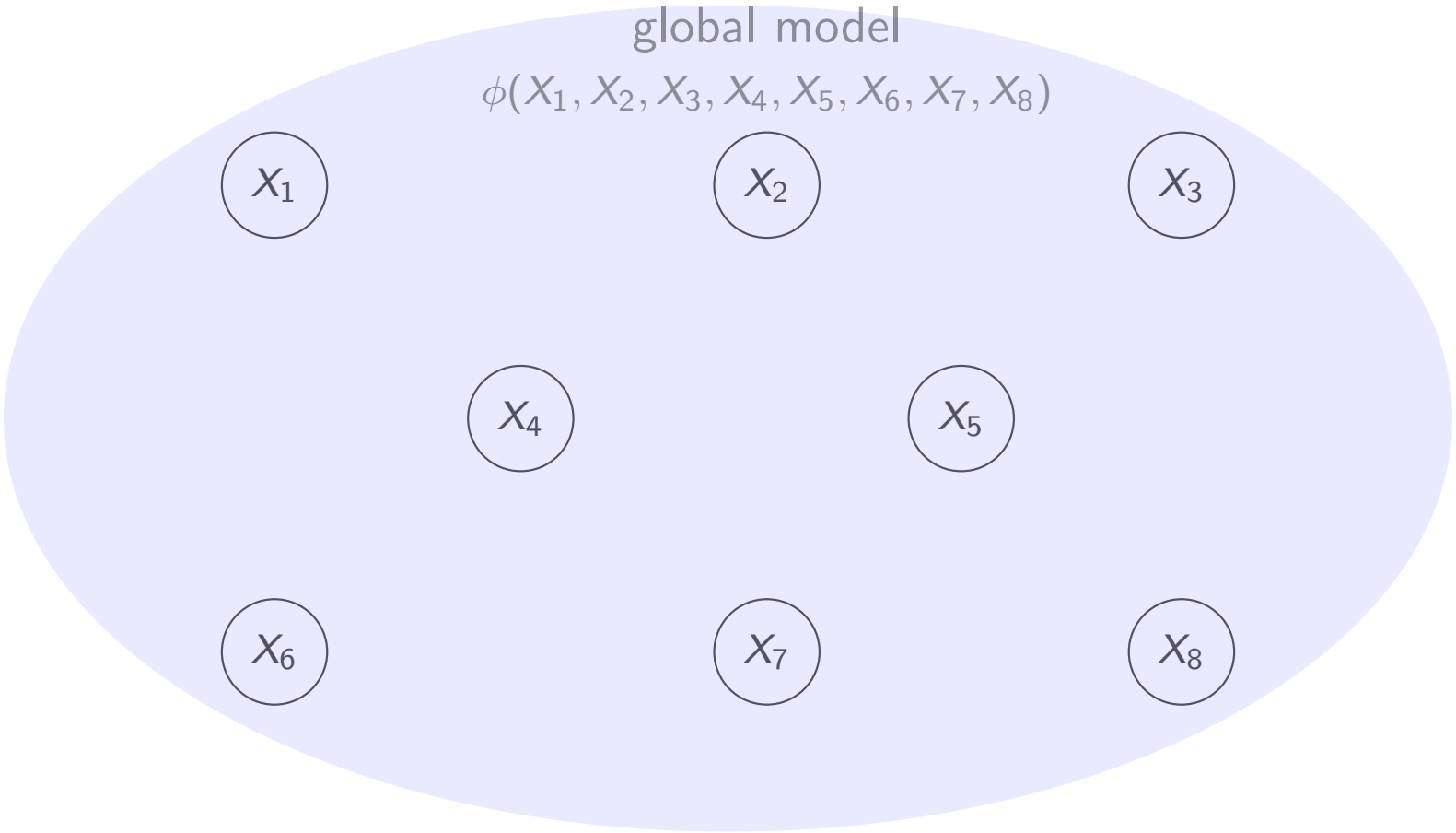
# Exercise

1. Compute $\underline{P}(r)$ by brute-force combinatorial approach
2. Approximate $\underline{P}(r)$ by linearizing the multilinear task

# Probabilistic Graphical Models

aka Decomposable Multivariate Probabilistic Models

(whose decomposability is induced by independence )

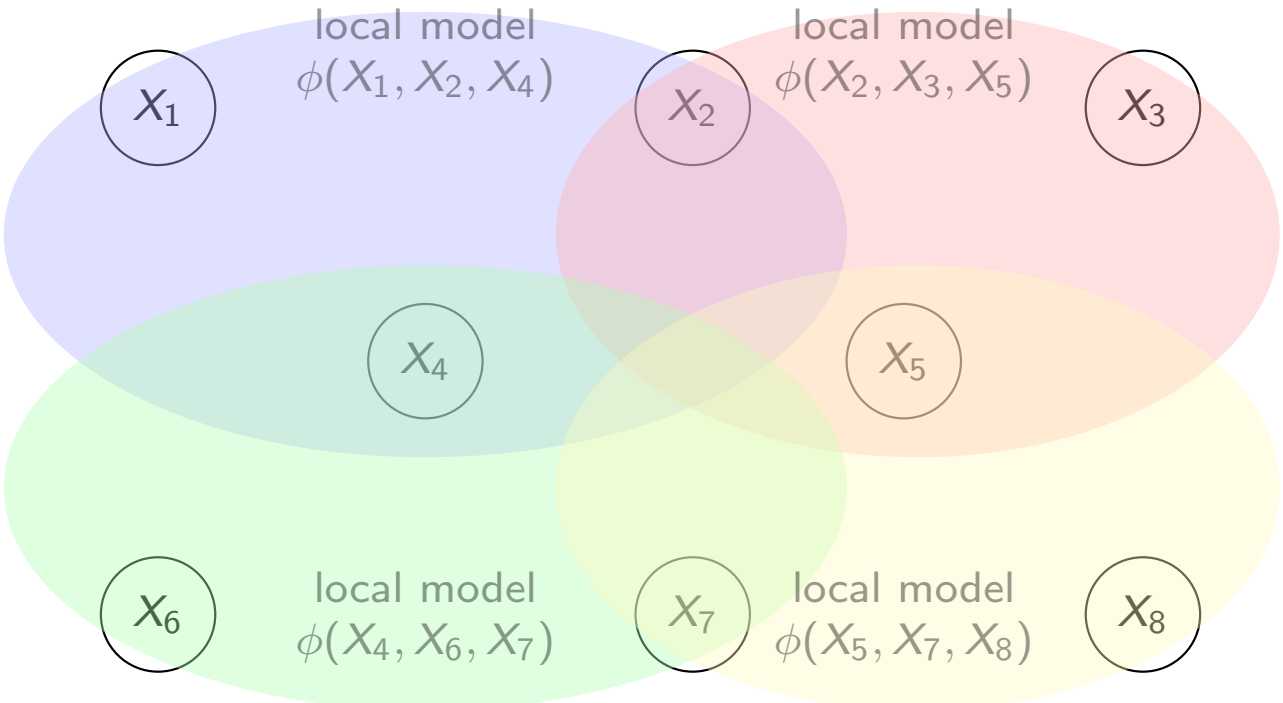


# Probabilistic Graphical Models

aka Decomposable Multivariate Probabilistic Models

(whose decomposability is induced by independence )

$$\phi(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \phi(X_1, X_2, X_4) \otimes \phi(X_2, X_3, X_5) \otimes \phi(X_4, X_6, X_7) \otimes \phi(X_5, X_7, X_8)$$



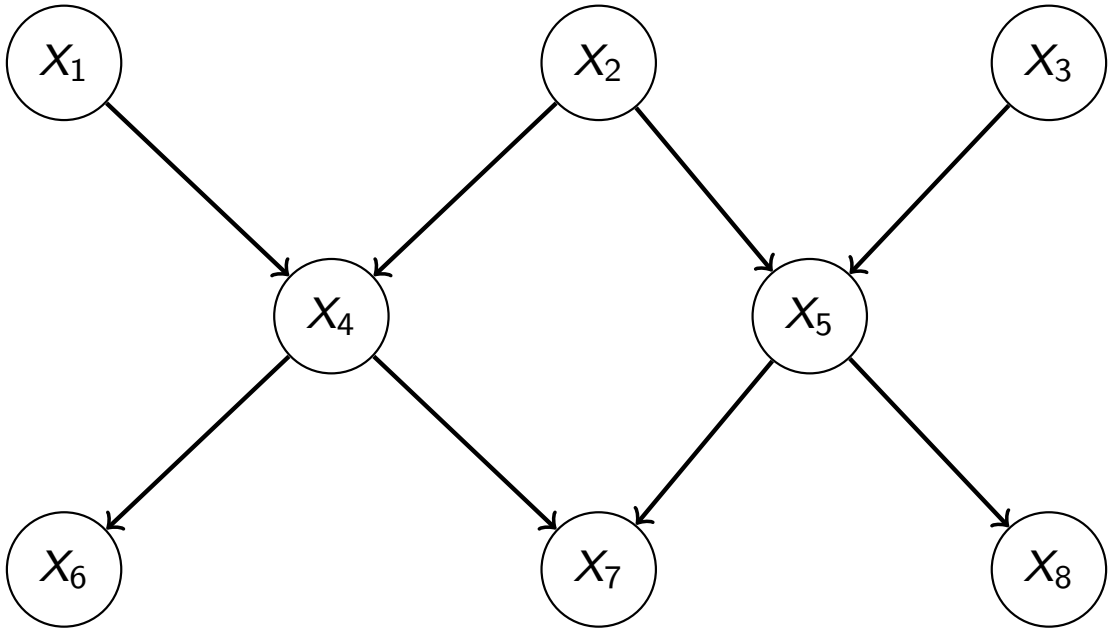
# Probabilistic Graphical Models

aka Decomposable Multivariate Probabilistic Models

(whose decomposability is induced by independence )

directed graphs

Bayesian/credal networks

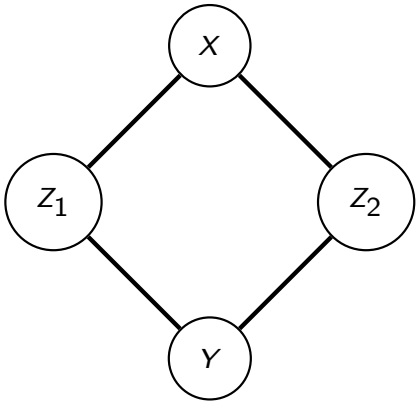


# Markov Condition

- Probabilistic model over set of variables $(X_1, \ldots, X_n)$ in one-to-one correspondence with the nodes of a graph

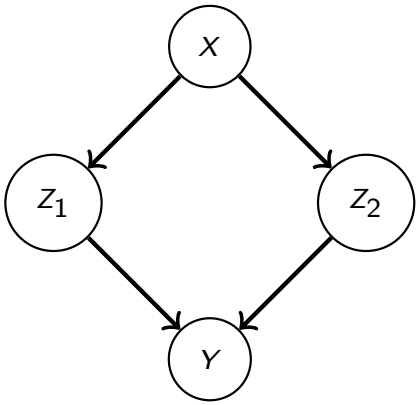
## Undirected Graphs

*$X$ and $Y$ are independent given $Z$ if any path between $X$ and $Y$ containts an element of $Z$*




## Directed Graphs

*Given its parents, every node is independent of its non-descendants non-parents*

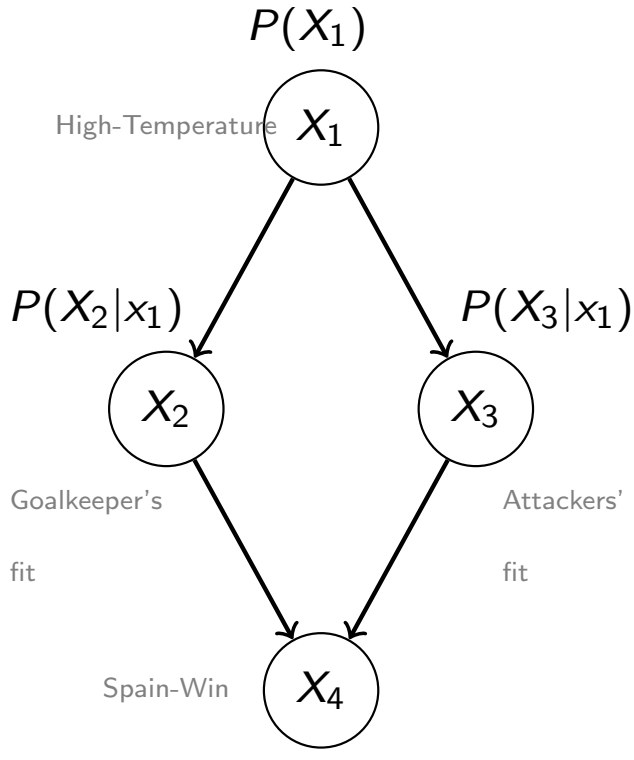



*$X$ and $Y$ are d-separated by $Z$ if, along every path between $X$ and $Y$ there is a $W$ such that either $W$ has converging arrows and is not in $Z$ and none of its descendants are in $Z$, or $W$ has no converging arrows and is in $Z$*

# Bayesian networks [Pearl, 1986]

- Set of categorical variables $X_1, \ldots, X_n$
- Directed acyclic graph
  - conditional (stochastic) independencies according to the Markov condition:

    "any node is conditionally independent of its non-descendents given its parents"
- A conditional mass function for each node and each possible value of the parents
  - $\{P(X_i|\mathrm{pa}(X_i)) \ , \forall i = 1, \ldots, n \ , \ \forall \mathrm{pa}(X_i) \ \}$
- Defines a joint probability mass function
  - $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\mathrm{pa}(X_i))$




$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_3, x_2)$$

*E.g., given temperature, fitnesses independent*

# Bayesian network - simple example - bn1.txt

```
library(bnlearn)
source('my.bn.inference.r')
net = model2network("[x1][x2|x1][x3|x1][x4|x2:x3]")

cpt1.x1 = matrix(c(0.7, 0.3), ncol = 2,
    dimnames = list(NULL, c('true', 'false')))
cpt1.x2 = c(0.1, 0.9, 0.3, 0.7)
dim(cpt1.x2) = c(2, 2)
dimnames(cpt1.x2) = list("x2" = c("true", "false"),
            "x1" =  c("true", "false"))
cpt1.x3 = c(0.5, 0.5, 0.2, 0.8)
dim(cpt1.x3) = c(2, 2)
dimnames(cpt1.x3) = list("x3" = c("true", "false"),
            "x1" =  c("true", "false"))
cpt1.x4 = c(0.9, 0.1, 0.5, 0.5, 0.4, 0.6, 0.1, 0.9)
dim(cpt1.x4) = c(2, 2, 2)
dimnames(cpt1.x4) = list("x4" = c("true", "false"),
            "x2" =  c("true", "false"),
            "x3" = c("true", "false"))
```

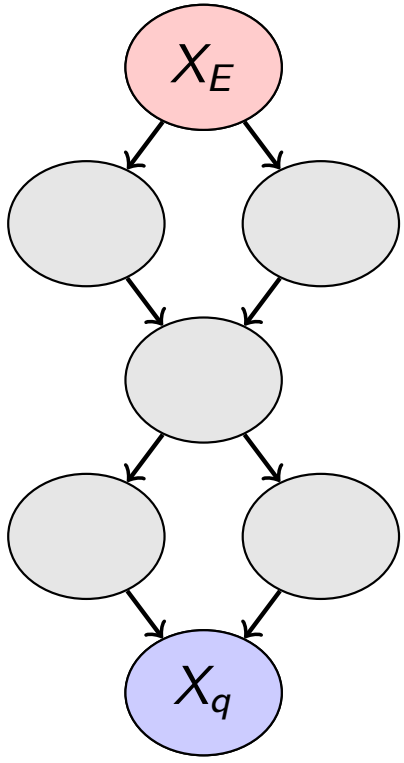
# Bayesian network - simple example - bn2.txt

```
net.1 = custom.fit(net, dist = list(x1=cpt1.x1,
                                    x2=cpt1.x2, x3=cpt1.x3, x4=cpt1.x4))

query=rep(NA,length(net.1))
names(query) <- names(net.1)
query[2]='false'
res <- my.bn.inference(net.1,query)

query[1]='true'
res <- my.bn.inference(net.1,query)

query[2]=NA
res <- my.bn.inference(net.1,query)

evidence=rep(NA,length(net.1))
names(evidence) <- names(net.1)
evidence[4]='true'
res <- my.bn.inference(net.1,query,evidence)

evidence[4]='false'
res <- my.bn.inference(net.1,query,evidence)
```

# Updating Bayesian networks

- Conditional probs for a variable of interest $X_q$ given observations $X_E = x_E$
- Updating Bayesian nets is NP-hard (fast algorithms for polytrees)

$$P(x_q|x_E) = \frac{P(x_q, x_E)}{P(x_E)} = \frac{\sum_{\mathbf{x}\setminus\{x_q,x_E\}} \prod_{i=1}^{n} P(x_i|\pi_i)}{\sum_{\mathbf{x}\setminus\{x_E\}} \prod_{i=1}^{n} P(x_i|\pi_i)}$$




$P(x_q|x_E) = .38$

# Credal networks [Cozman, 2000]

- Generalization of BNs to imprecise probabilities
- Credal sets instead of prob mass functions

  $\{P(X_i|\mathrm{pa}(X_i))\} \Rightarrow \{K(X_i|\mathrm{pa}(X_i))\}$
- Strong (instead of stochastic) independence in the semantics of the Markov condition (We will talk about credal nets with strong independence, because it has been around for more time, so we have more applications for it.)
- Convex set of joint mass functions

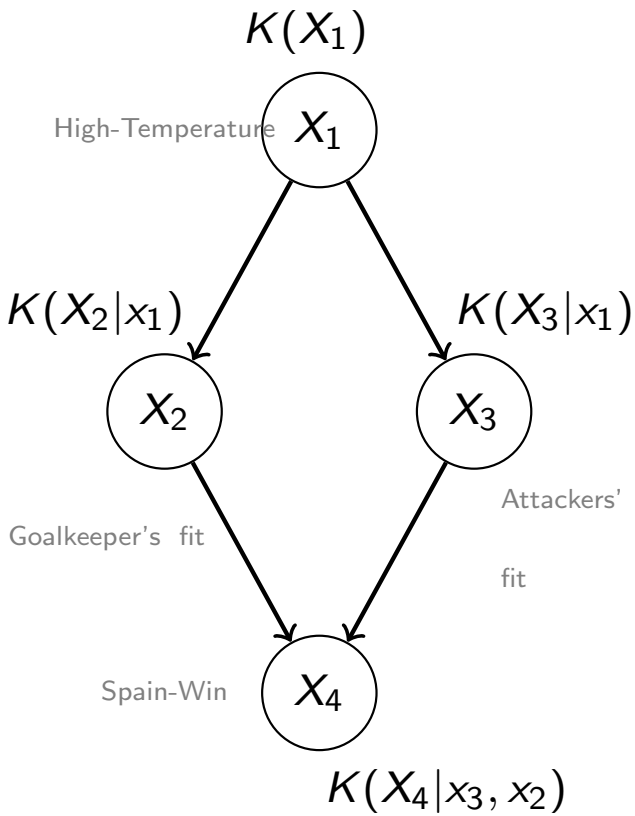
  $K(X_1,\ldots,X_n) = \mathrm{CH}\Big\{P(X_1,\ldots,X_n)\Big\}$

  $P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i|\mathrm{pa}(X_i))$

  $\forall P(X_i|\mathrm{pa}(X_i)) \in K(X_i|\mathrm{pa}(X_i))$

  $\forall i = 1,\ldots,n \quad \forall \mathrm{pa}(X_i)$
- Every conditional mass function takes values in its credal set independently of the others

  CN $\equiv$ (exponential) number of BNs




*E.g., $K(X_1)$ defined by constraint $P(x_1) > .7$, very likely to be warm*

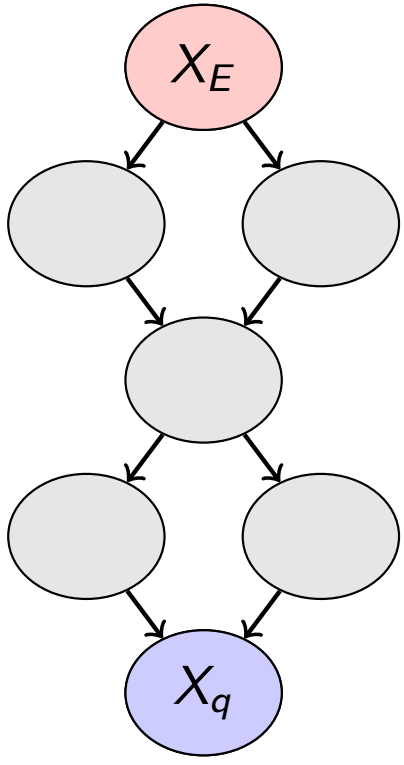
# Credal network - simple example - cn1.txt

```
source('my.cn.inference.r')
cpt2.x1 = matrix(c(1, 0), ncol = 2,
    dimnames = list(NULL, c('true', 'false')))

# In this part of the talk, we use a very simplistic
# representation of binary credal networks:
#
# Two BNs, each one gives one of the vertices of
# each local credal set

net.2 = custom.fit(net, dist = list(x1=cpt2.x1,
                x2=cpt1.x2, x3=cpt1.x3, x4=cpt1.x4))

# So net.2 is precise apart from variable x1, which
# has 0.7 <= P(x1) <= 1
```

# Updating credal networks

- Conditional probs for a variable of interest $X_q$ given observations $X_E = x_E$
- Updating Bayesian nets is NP-hard
  (fast algorithms for polytrees)

$$P(x_q|x_E) = \frac{P(x_q, x_E)}{P(x_E)} = \frac{\sum_{\mathbf{x}\setminus\{x_q,x_E\}} \prod_{i=1}^{n} P(x_i|\pi_i)}{\sum_{\mathbf{x}\setminus\{x_E\}} \prod_{i=1}^{n} P(x_i|\pi_i)}$$




- Updating credal nets is $\mathrm{NP}^{\mathrm{PP}}$-hard,
  NP-hard on polytrees [Mauá et al.]
  Easy in trees under epistemic irr. [de Cooman et al.]

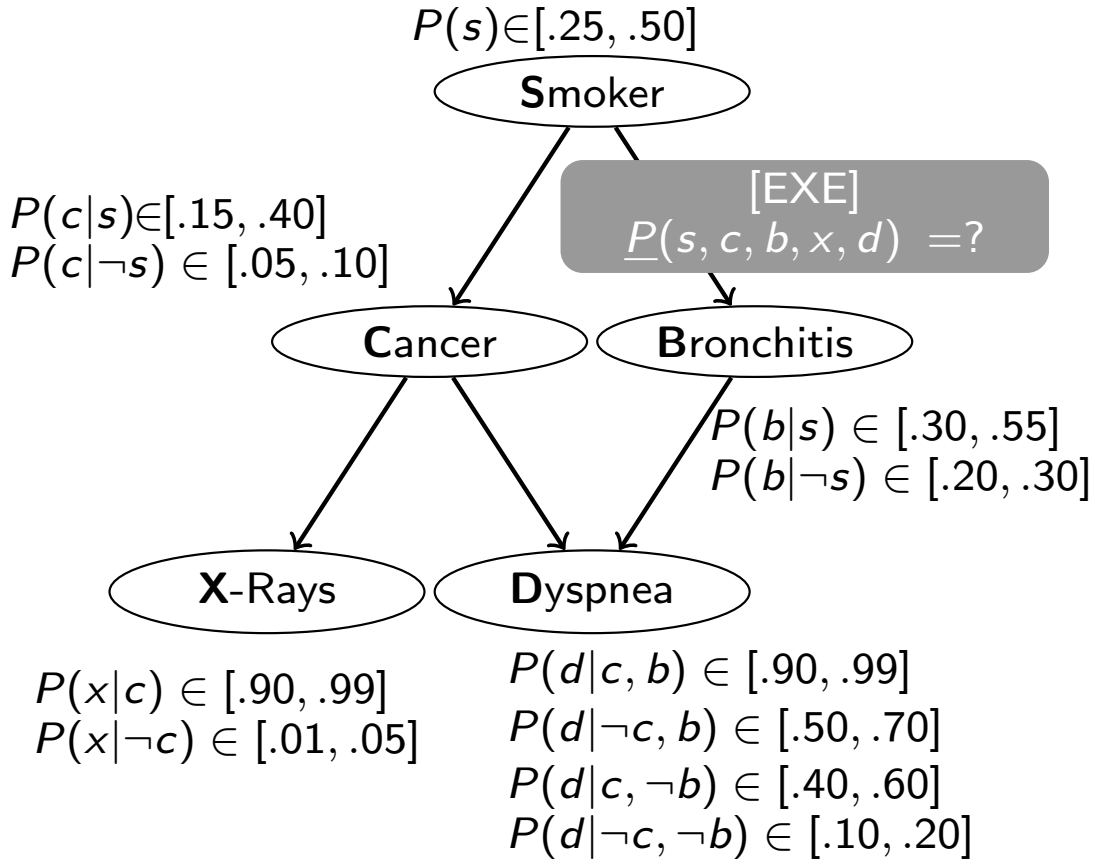
$$\underline{P}(x_q|x_E) = \min_{\substack{P(X_i|\pi_i)\in K(X_i|\pi_i) \\ i=1,\ldots,n}} \frac{\sum_{\mathbf{x}\setminus\{x_Q,x_E\}} \prod_{i=1}^{n} P(x_i|\pi_i)}{\sum_{\mathbf{x}\setminus\{x_Q\}} \prod_{i=1}^{n} P(x_i|\pi_i)}$$

# Credal network - simple updating example - cn2.txt

```
source('my.cn.inference.r')
cpt2.x1 = matrix(c(1, 0), ncol = 2,
    dimnames = list(NULL, c('true', 'false')))

net.2 = custom.fit(net, dist = list(x1=cpt2.x1,
                x2=cpt1.x2, x3=cpt1.x3, x4=cpt1.x4))

query=rep(NA,length(net.1))
names(query) <- names(net.1)
query[2]='false'
res <- my.cn.inference(net.1,net.2,query)

query[1]='true'
res <- my.cn.inference(net.1,net.2,query)
cat('Query p(',res$query,'|',res$evidence,
    ') -- interval result: [',res$min.p,
    ',',res$max.p,']\n')
```

# Medical diagnosis by CNs (a simple example of)

- Five Boolean vars
- Conditional independence relations given by a DAG
- Elicitation of the local (conditional) CSs
- This is a CN specification
- The strong extension $K(S, C, B, X, D) =$




$$\mathrm{CH}\left\{P(S, C, B, X, D) \;\middle|\; \begin{array}{l} P(s, c, b, x, d)=P(s)P(c|s)P(b|s)P(x|c)P(d|c, b) \\ P(S) \in K(S) \\ P(C|s) \in K(C|s), P(C|\neg s) \in K(C|\neg s) \\ \ldots \end{array}\right\}$$

# Asia network example - simpleasia1.txt

```
library(bnlearn)

net = model2network("[smoke][lung|smoke][bronc|smoke][xrays|lung][dysp|lung:bronc]")

cpt1.smoke = matrix(c(0.25, 0.75), ncol = 2,
   dimnames = list(NULL, c('true', 'false')))
cpt2.smoke = matrix(c(0.5, 0.5), ncol = 2,
   dimnames = list(NULL, c('true', 'false')))

cpt1.lung = c(0.15, 0.85, 0.05, 0.95)
dim(cpt1.lung) = c(2, 2)
dimnames(cpt1.lung) = list("lung" = c("true", "false"),
"smoke" =  c("true", "false"))
cpt2.lung = c(0.4, 0.6, 0.1, 0.9)
dim(cpt2.lung) = c(2, 2)
dimnames(cpt2.lung) = list("lung" = c("true", "false"),
"smoke" =  c("true", "false"))
cpt1.bronc = c(0.3, 0.7, 0.2, 0.8)
dim(cpt1.bronc) = c(2, 2)
dimnames(cpt1.bronc) = list("bronc" = c("true", "false"),
 "smoke" =  c("true", "false"))
cpt2.bronc = c(0.55, 0.45, 0.3, 0.7)
dim(cpt2.bronc) = c(2, 2)
dimnames(cpt2.bronc) = list("bronc" = c("true", "false"),
 "smoke" =  c("true", "false"))
cpt1.xrays = c(0.9, 0.1, 0.01, 0.99)
dim(cpt1.xrays) = c(2, 2)
dimnames(cpt1.xrays) = list("xrays" = c("true", "false"),
 "lung" =  c("true", "false"))
cpt2.xrays = c(0.99, 0.01, 0.05, 0.95)
dim(cpt2.xrays) = c(2, 2)
dimnames(cpt2.xrays) = list("xrays" = c("true", "false"),
 "lung" =  c("true", "false"))
```

# Asia network example - simpleasia2.txt

```
cpt1.dysp = c(0.9, 0.1, 0.5, 0.5, 0.4, 0.6, 0.1, 0.9)
dim(cpt1.dysp) = c(2, 2, 2)
dimnames(cpt1.dysp) = list("dysp" = c("true", "false"),
"lung" =  c("true", "false"), "bronc" = c("true", "false"))
cpt2.dysp = c(0.99, 0.01, 0.7, 0.3, 0.6, 0.4, 0.2, 0.8)
dim(cpt2.dysp) = c(2, 2, 2)
dimnames(cpt2.dysp) = list("dysp" = c("true", "false"),
"lung" =  c("true", "false"), "bronc" = c("true", "false"))
net.1 = custom.fit(net, dist = list(smoke=cpt1.smoke,
  lung=cpt1.lung, bronc=cpt1.bronc, xrays=cpt1.xrays, dysp=cpt1.dysp))
net.2 = custom.fit(net, dist = list(smoke=cpt2.smoke,
  lung=cpt2.lung, bronc=cpt2.bronc, xrays=cpt2.xrays, dysp=cpt2.dysp))
query=rep('true',length(net.1))
names(query) <- names(net.1)
source('my.cn.inference.r')
res <- my.cn.inference(net.1,net.2,query)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')
```

# Asia network example - simpleasia3.txt

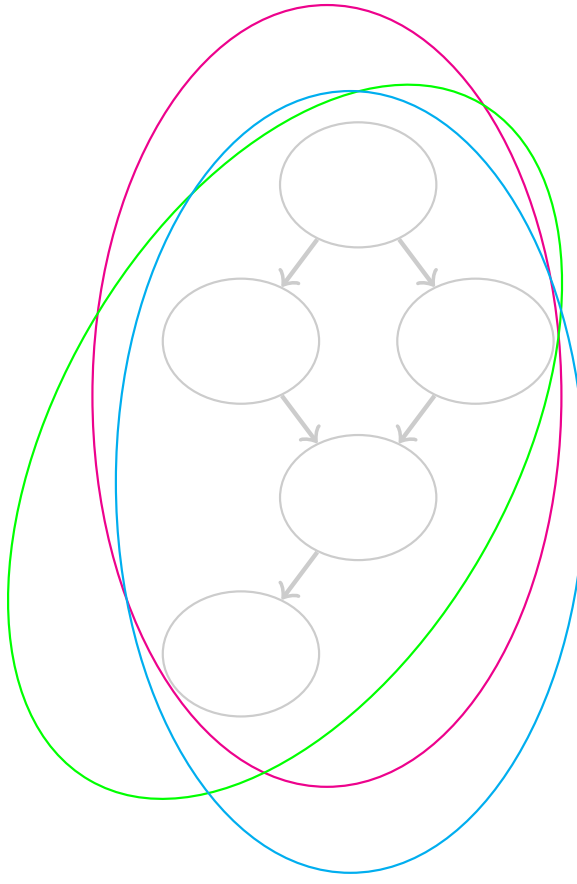
```
query=rep(NA,length(net.1))
names(query) <- names(net.1)
query['lung'] <- 'true'

evi=rep(NA,length(net.1))
names(evi) <- names(net.1)
evi['dysp'] <- 'true'
res <- my.cn.inference(net.1,net.2,query,evi)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')

evi['bronc'] <- 'true'
res <- my.cn.inference(net.1,net.2,query,evi)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')

evi['bronc'] <- 'false'
evi['xrays'] <- 'true'
res <- my.cn.inference(net.1,net.2,query,evi)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')
```

# Asia network example - simpleasia4.txt

```
evi=rep(NA,length(net.1))
names(evi) <- names(net.1)

evi['bronc'] <- 'false'
res <- my.cn.inference(net.1,net.2,query,evi)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')
evi['smoke'] <- 'true'
res <- my.cn.inference(net.1,net.2,query,evi)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')
evi['bronc'] <- 'true'
res <- my.cn.inference(net.1,net.2,query,evi)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')
evi['dysp'] <- 'true'
res <- my.cn.inference(net.1,net.2,query,evi)
cat('Query p(',res$query,'|',res$evidence,
   ') -- interval result: [',res$min.p,',',res$max.p,']\n')
```

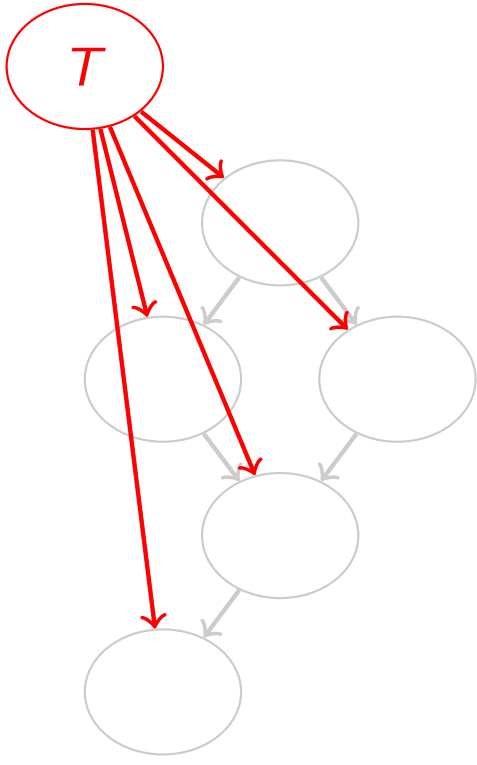
# Non-separately specified CNs

- Constraints among different conditional mass functions of a CN
- Explicit enumeration of the relative BNs
  - Auxiliary parent selecting the conditional probabilities [Cano, Cano, Moral, 1994] with a vacuous prior
- "Extensive" specification
  - Constraints among conditional mass functions of the same variable
  - Each CPT takes values from a set of tables an auxiliary parent selecting the tables



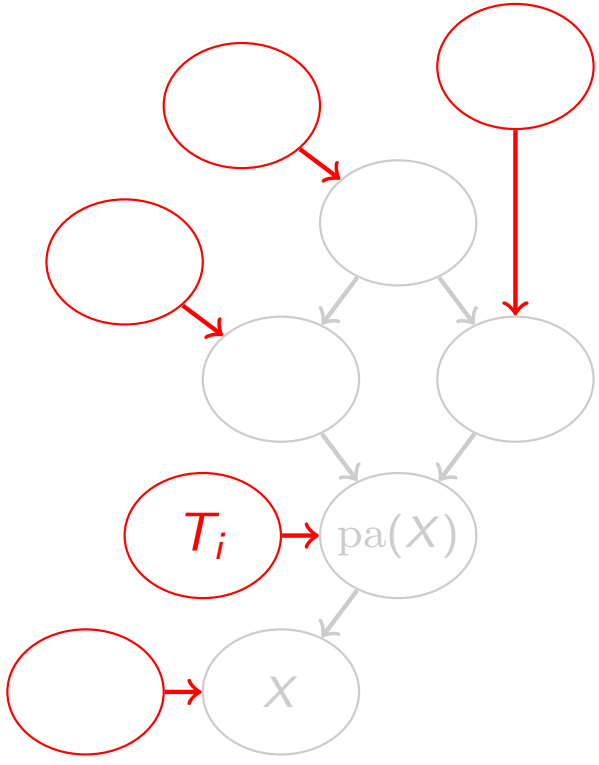
# Non-separately specified CNs

- Constraints among different conditional mass functions of a CN
- Explicit enumeration of the relative BNs
  - Auxiliary parent selecting the conditional probabilities [Cano, Cano, Moral, 1994] with a vacuous prior
- "Extensive" specification
  - Constraints among conditional mass functions of the same variable
  - Each CPT takes values from a set of tables an auxiliary parent selecting the tables



# Non-separately specified CNs

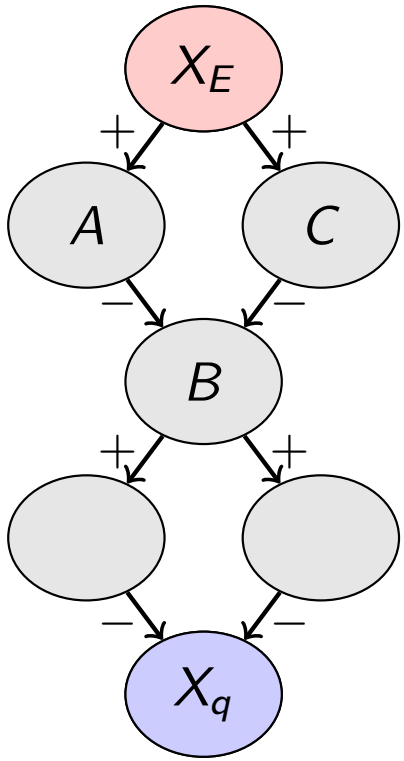
- Constraints among different conditional mass functions of a CN
- Explicit enumeration of the relative BNs
  - Auxiliary parent selecting the conditional probabilities [Cano, Cano, Moral, 1994] with a vacuous prior
- "Extensive" specification
  - Constraints among conditional mass functions of the same variable
  - Each CPT takes values from a set of tables an auxiliary parent selecting the tables




$$P(X|\mathrm{pa}(X),\ T = t_j\ ) = P_j(X|\mathrm{pa}(X))$$

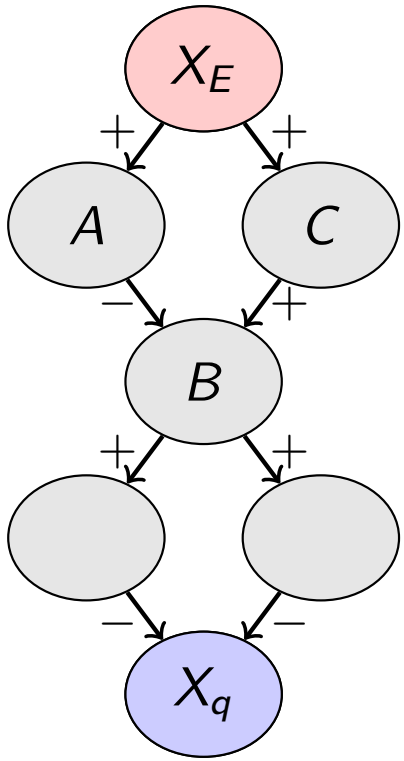
# Qualitative Probabilistic Networks

- No numerical specifications.
- Qualitative relations: $A$ influences $B$ positively if $P(b|a, C) \geq P(b|\neg a, C)$ for every $C$. Each arc is marked with a sign: $+$, $-$, 0 or ?.
- Query: does it hold $P(x_q|x_E) \geq P(x_q)$ for every $P$?
- It can be solved by sign propagation ideas. The basic idea is that influences are "transitive".
- $\forall P: \ P(x_q|x_E) \geq P(x_q)$? Yes, it is true!
- $\forall P: \ P(x_q|x_E) \geq P(x_q)$? Now it is not!



# Qualitative Probabilistic Networks

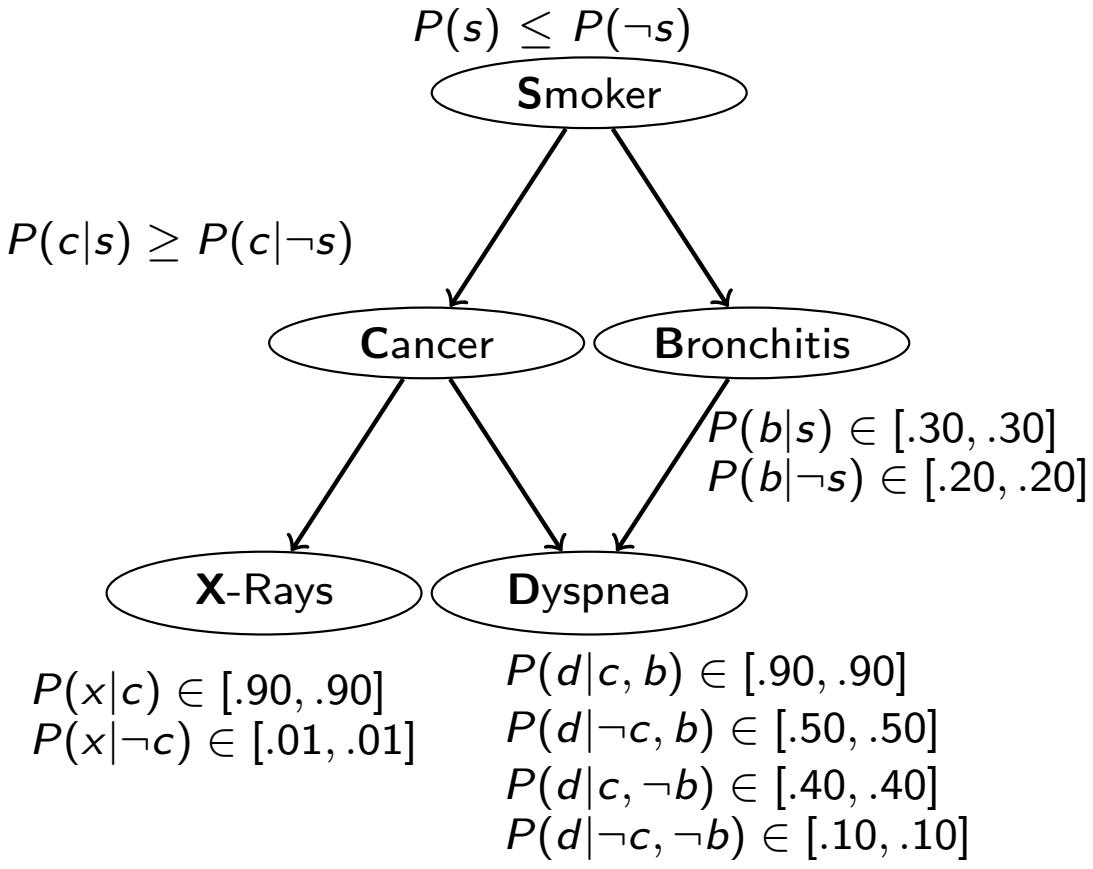
- No numerical specifications.
- Qualitative relations: $A$ influences $B$ positively if $P(b|a, C) \geq P(b|\neg a, C)$ for every $C$. Each arc is marked with a sign: $+$, $-$, 0 or ?.
- Query: does it hold $P(x_q|x_E) \geq P(x_q)$ for every $P$?
- It can be solved by sign propagation ideas. The basic idea is that influences are "transitive".
- $\forall P: \ P(x_q|x_E) \geq P(x_q)$? Yes, it is true!
- $\forall P: \ P(x_q|x_E) \geq P(x_q)$? Now it is not!



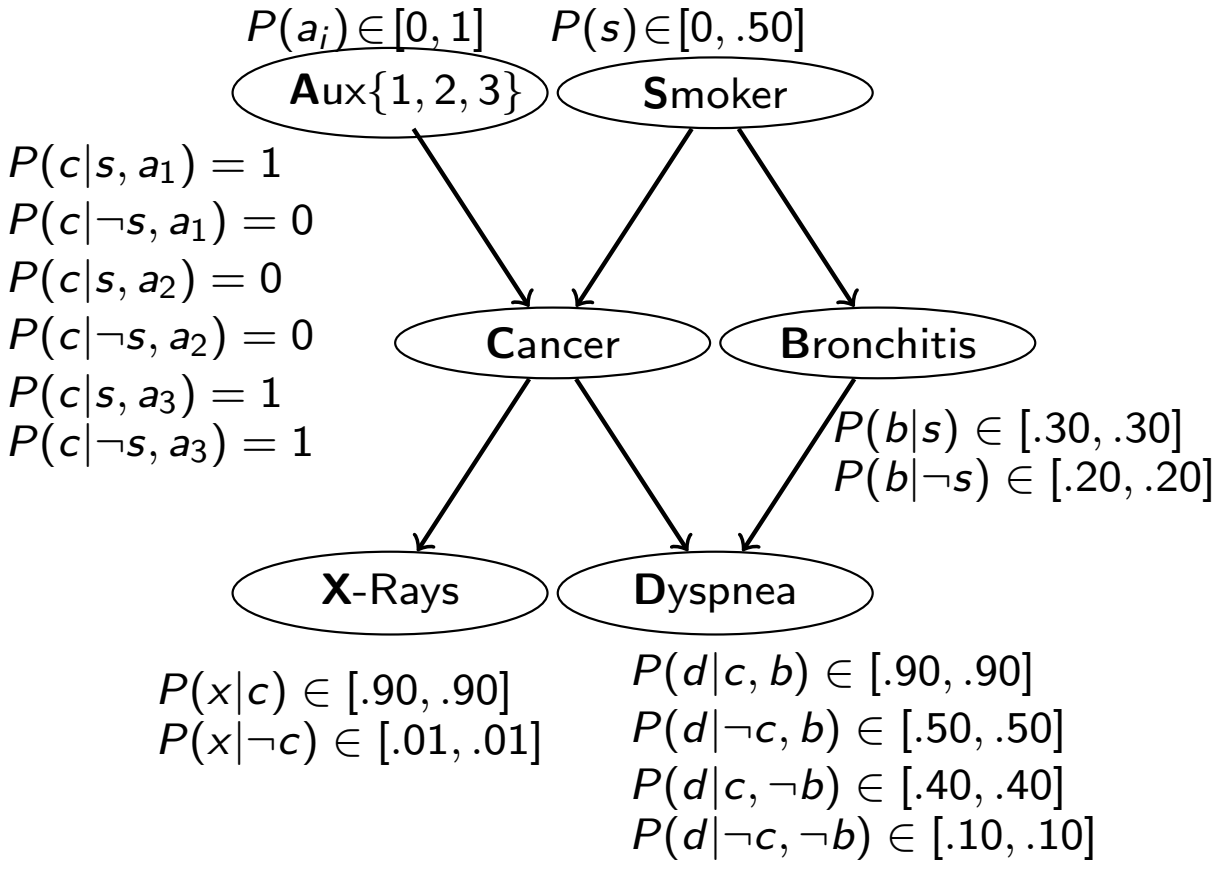
# Qualitative Probabilistic Networks

- QPNs are a simple type of credal networks.
- QPNs: more "uncertainty" does not mean inferences are harder!
- Mix of QPNs with numerical parameters becomes the Semi-QPNs. They are as powerful as general credal networks.
- Learning the parameters of a Bayesian networks can be seen as a certain optimization over QPNs/credal networks.

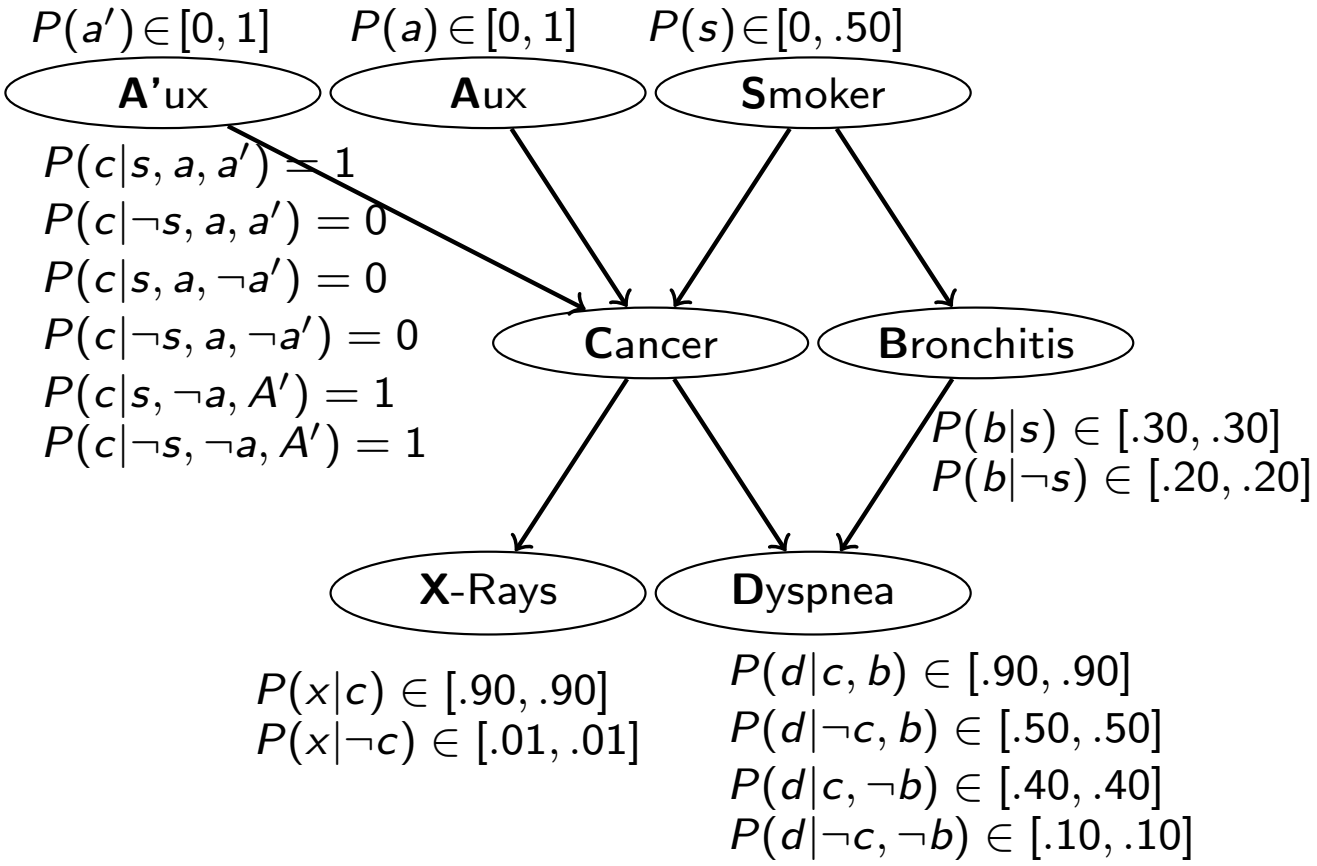
# Semi-Qualitative Probabilistic Networks



# Semi-Qualitative Probabilistic Networks



# Semi-Qualitative Probabilistic Networks



# Semi-Qualitative Probabilistic Networks

- Can we represent this network using the same software as before? Yes!
- Can we compute inferences using the software as before?
- $\overline{P}(c|b,x)$ is obtained by a fractional linear optimization.
- $\max_P(P(c|b) - P(c)) > 0 \iff \max_P(P(c,b) - P(c) \cdot P(b)) > 0$ (assuming $P(b) > 0$) is a non-linear query.
- Non-linear inference might need additional effort.

# Motivations

- Hypothesis tests are ubiquitous. Decision making, scientific discoveries, feature selection in many fields (e.g., medical, demographic, environmental, etc.) are based on the results of hypothesis tests.
- In case of scarce prior knowledge of the distributions of interest, **nonparametric tests** are preferred (robust to outliers, do not assume normality of the data,...)
- Frequentist procedures (null hypothesis significance test) are usually adopted. However
  - Frequentist hypothesis tests cannot assess evidence for the null hypothesis.
  - Lack of a sound criterion for deciding Type I error (usually 0.05 or 0.01).
  - the p-value and thus the outcome of the test depend on the intention of the person who has collected the data.

# Motivations

- A Bayesian nonparametric approach evaluates the posterior probability of the alternative hypotheses;
- This allows taking decisions which minimize the expected loss once the costs of type I and type II errors are specified.
- The outcome of the test depends only on the prior belief and on the collected data.

**Why should we leave nonparametric hypothesis tests to frequentist procedures?**

# The Dirichlet Process

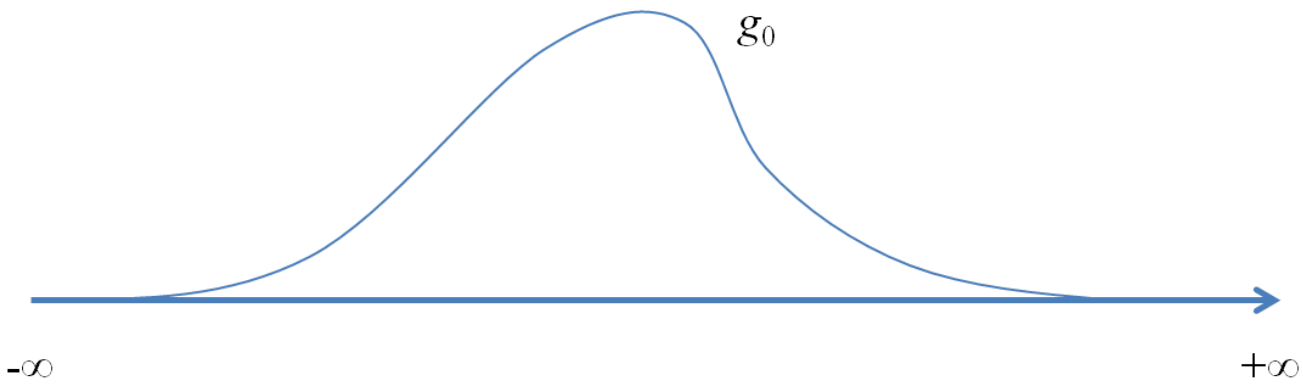
- The Dirichlet process (DP) is one of the most popular Bayesian nonparametric models.
- In Ferguson's seminal paper it is used to estimate:
  - distribution function, mean, quantiles, variance;
  - $P(X \leq Y) \rightarrow$ **Mann-Whitney statistic**;
  - *survival function* $\rightarrow$ **Kaplan-Meier estimator** (Susarla, Van Ryzin and Blum);
  - measure of *bivariate dependence* $\rightarrow$ **Kendall's tau** (Dalal and Phadia);
  - ...
- DP provides Bayesian justification of traditional nonparametric estimators.
- Can be used to perform traditional tests in a Bayesian way.

# The Dirichlet Process - definition

- The DP is a way of assigning a probability distribution over probability distributions.
- Let the probability measure $P$ be the Dirichlet process $\mathrm{Dp}(s, g_0)$:
  - $s =$ prior strength (scalar);
  - $g_0 =$ prior base probability measure on $\mathbb{X}$ (infinite dimensions).
- Then, given a finite partition $B_1, B_2, \ldots, B_m$ of $\mathbb{X}$, $(P(B_1), \ldots, P(B_m)) \sim \mathrm{Dir}(s\ g_0(B_1), \ldots, s\ g_0(B_m))$.

# The Dirichlet Process on $\mathbb{R}$

- Sample space $\mathbb{X} = \mathbb{R}$;
- $P \sim \mathrm{Dp}(s, g_0)$;

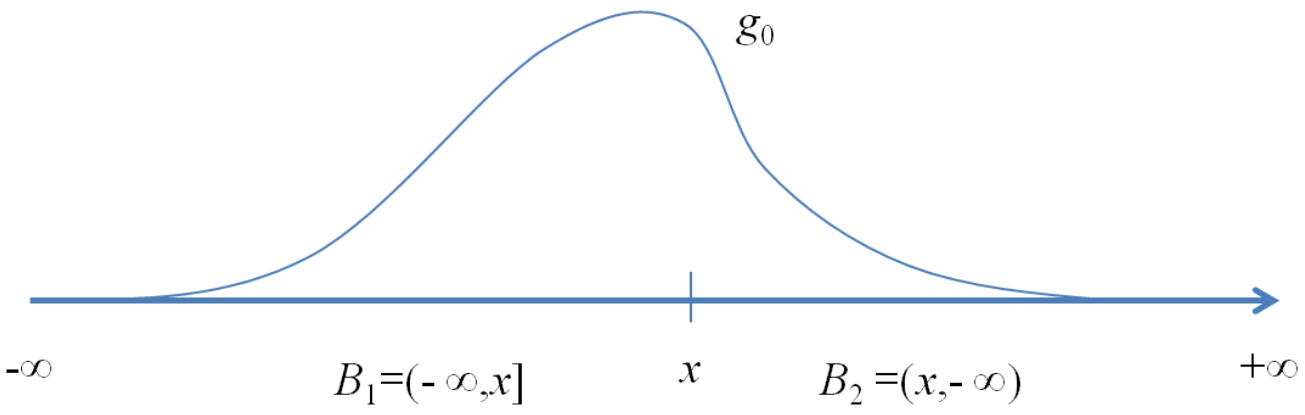



$$P(X < x) = P(B_1) \sim \mathrm{Dir}(sg_0(B_1), sg_0(B_2))$$
$$\sim \mathrm{Beta}(sg_0(B_1), s[1 - g_0(B_1)]) \implies \mathcal{E}[P(X < x)] = g_0$$

**$g_0$ represents our prior belief about the shape of $P$.**

# The Dirichlet Process on $\mathbb{R}$

- Sample space $\mathbb{X} = \mathbb{R}$;
- $P \sim \mathrm{Dp}(s, g_0)$;

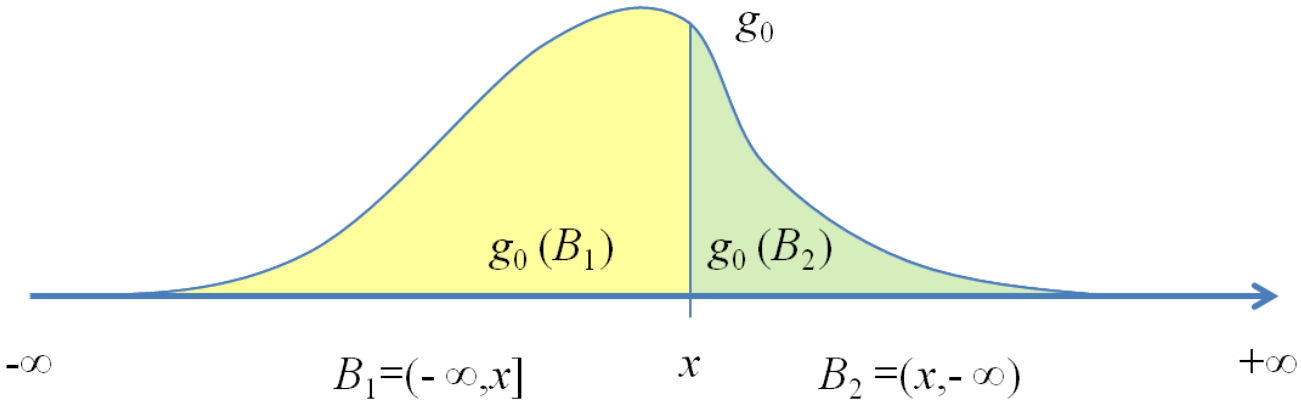



$$P(X < x) = P(B_1) \sim \mathrm{Dir}(sg_0(B_1), sg_0(B_2))$$
$$\sim \mathrm{Beta}(sg_0(B_1), s[1 - g_0(B_1)]) \implies \mathcal{E}[P(X < x)] = g_0$$

**$g_0$ represents our prior belief about the shape of $P$.**

# The Dirichlet Process on $\mathbb{R}$

- Sample space $\mathbb{X} = \mathbb{R}$;
- $P \sim \mathrm{Dp}(s, g_0)$;

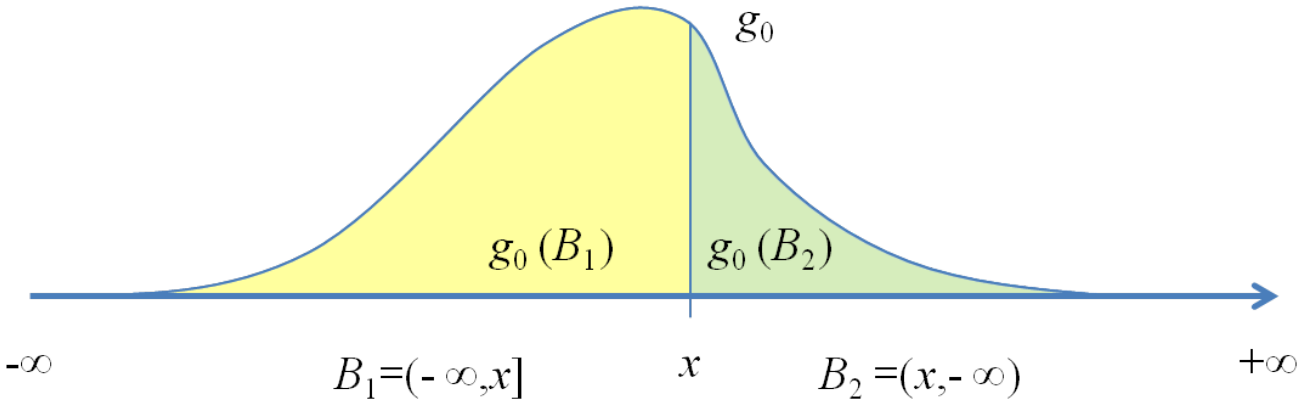



$$P(X < x) = P(B_1) \sim \mathrm{Dir}(sg_0(B_1), sg_0(B_2))$$
$$\sim \mathrm{Beta}(sg_0(B_1), s[1 - g_0(B_1)]) \implies \mathcal{E}[P(X < x)] = g_0$$

$g_0$ represents our prior belief about the shape of $P$.

# The Dirichlet Process on $\mathbb{R}$

- Sample space $\mathbb{X} = \mathbb{R}$;
- $P \sim \mathrm{Dp}(s, g_0)$;

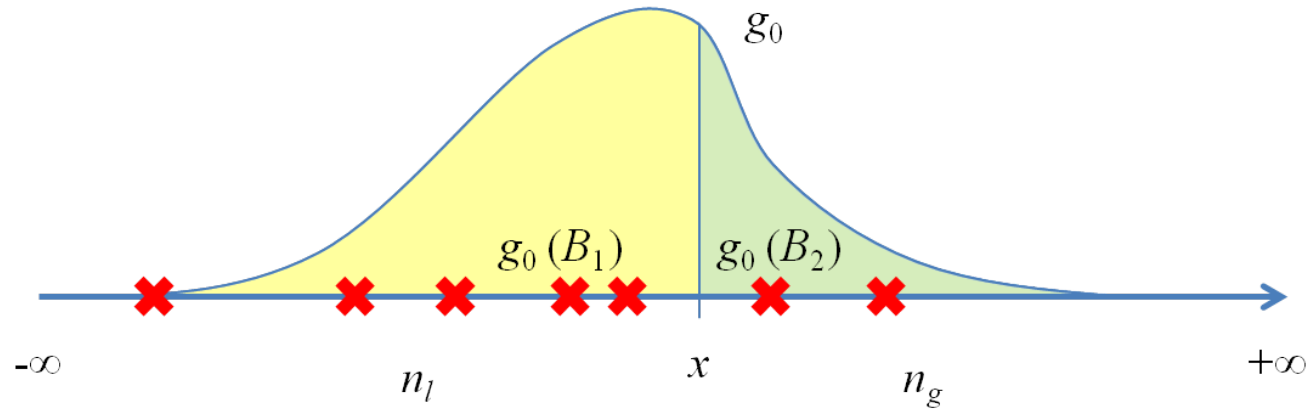



$$P(X < x) = P(B_1) \sim \mathrm{Dir}(sg_0(B_1), sg_0(B_2))$$
$$\sim \mathrm{Beta}(sg_0(B_1), s[1 - g_0(B_1)]) \implies \mathcal{E}[P(X < x)] = g_0$$

**$g_0$ represents our prior belief about the shape of $P$.**

# The Dirichlet Process on $\mathbb{R}$

Let $X_1, X_2, \ldots$ be $n$ samples from $P$




Prior: $P(X < x) \sim \text{Dir}(sg_0(B_1), sg_0(B_2))$

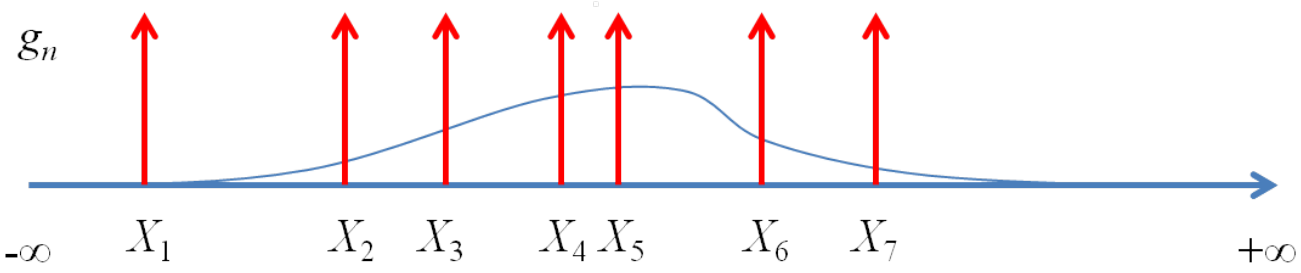
Posterior: $P(X < x) \sim \text{Dir}(sg_0(B_1) + n_l, sg_0(B_2) + n_g)$

$$\implies \mathcal{E}[P(X < x)] = \frac{sg_0(B_1) + n_l}{s + n}$$

# The Dirichlet Process on $\mathbb{R}$ - Conjugacy

Prior: $P \sim \mathrm{Dp}(s, g_0)$

Posterior: $P \sim \mathrm{Dp}(s+n, \underbrace{\frac{s}{s+n} g_0 + \frac{1}{n+s} \sum_{i=1}^{n} \delta_{X_i}}_{g_n})$



# Prior elicitation

In Bayesian analysis the problem is how to select $s$ and $g_0$.

Note that $g_0$ is a probability measure and, thus, very detailed information is needed for its elicitation.

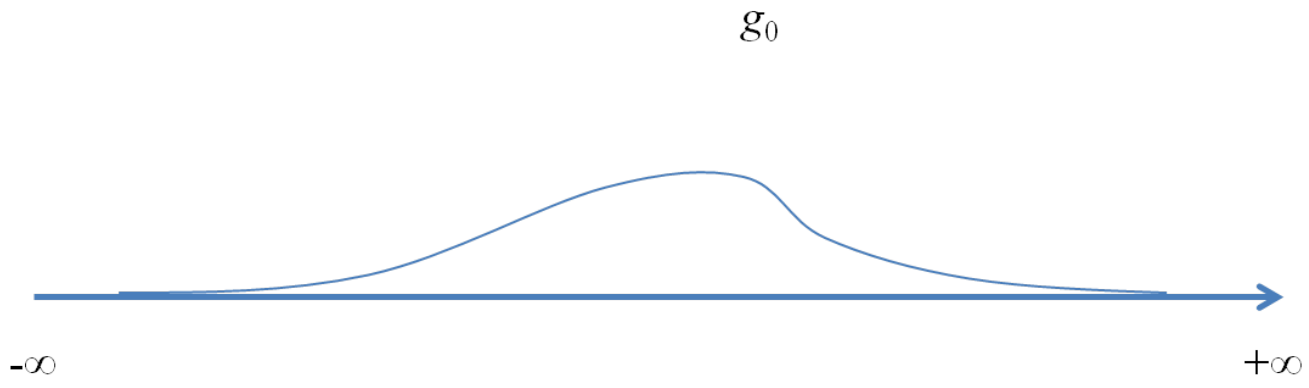
In case of lack of prior information:

- $s \to 0$ (Ferguson, Rubin);
- $s, G_0$ are selected using empirical Bayesian approaches;
- hierarchical prior on $s, G_0$.

# Near-ignorance solution: Imprecise Dirichlet Process

Keep $s$ fixed and let $g_0$ vary in the set of all probability measures:

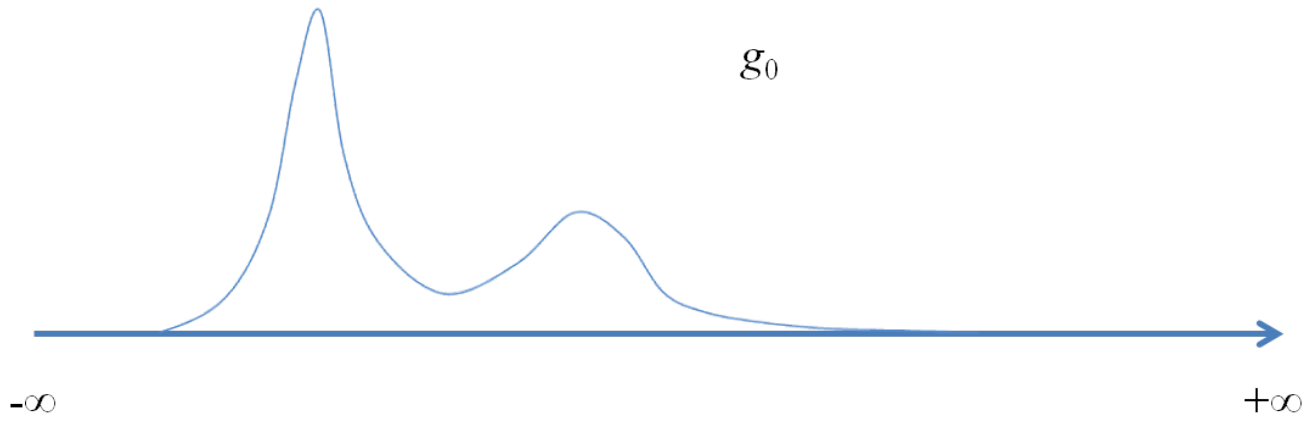
$$IDP: \quad \{Dp(s, g_0), \quad g_0 \in \mathbb{P}\}$$



# Near-ignorance solution: Imprecise Dirichlet Process

Keep $s$ fixed and let $g_0$ vary in the set of all probability measures:

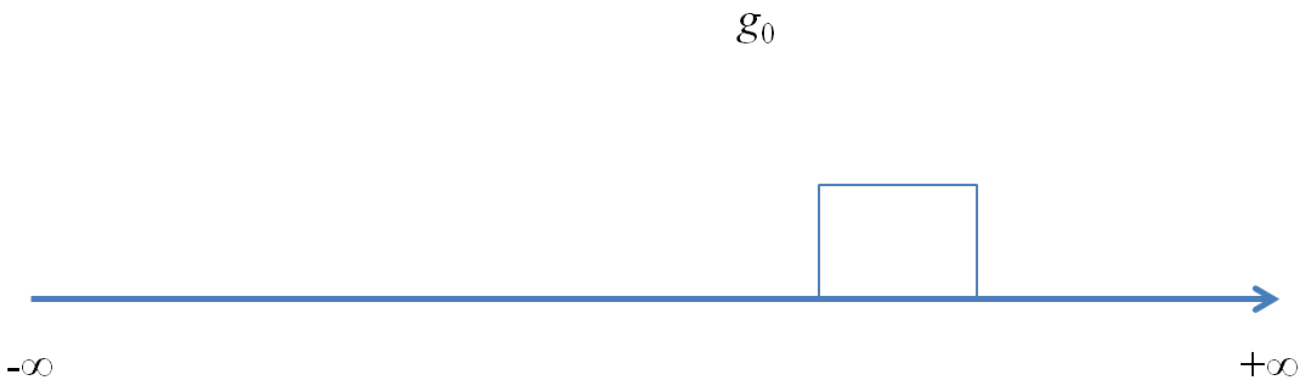
$$IDP: \quad \{Dp(s, g_0), \quad g_0 \in \mathbb{P}\}$$



# Near-ignorance solution: Imprecise Dirichlet Process

Keep $s$ fixed and let $g_0$ vary in the set of all probability measures:

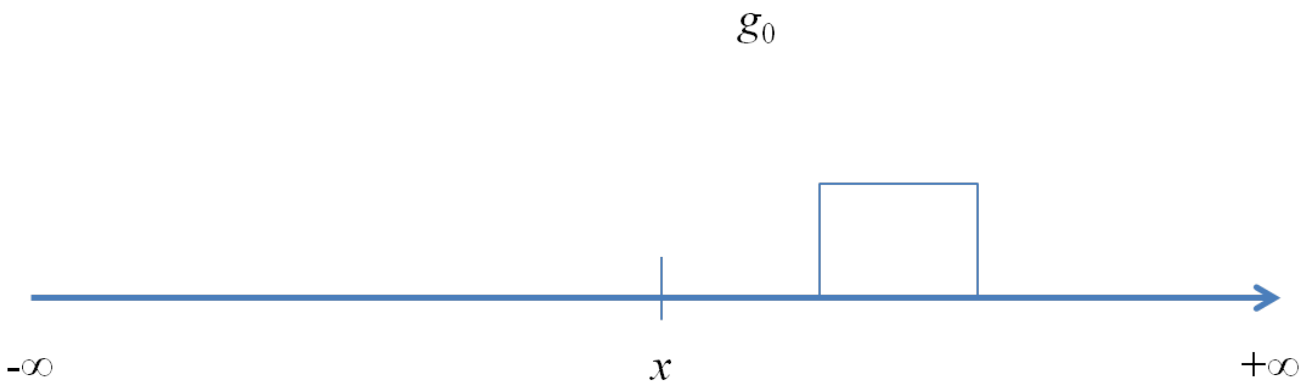
$$IDP: \quad \{Dp(s, g_0), \quad g_0 \in \mathbb{P}\}$$



# Near-ignorance solution: Imprecise Dirichlet Process

Keep $s$ fixed and let $g_0$ vary in the set of all probability measures:

$$IDP: \quad \{Dp(s, g_0), \quad g_0 \in \mathbb{P}\}$$

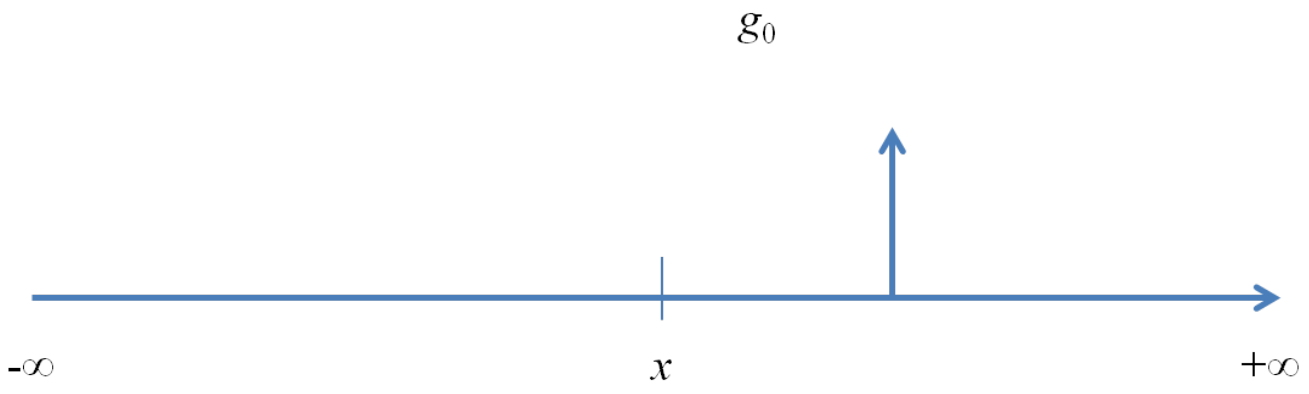



$$\mathcal{E}[P(X < x)] = g_0(-\infty, x] = 0$$

# Near-ignorance solution: Imprecise Dirichlet Process

Keep $s$ fixed and let $g_0$ vary in the set of all probability measures:

$$IDP: \quad \{Dp(s, g_0), \quad g_0 \in \mathbb{P}\}$$

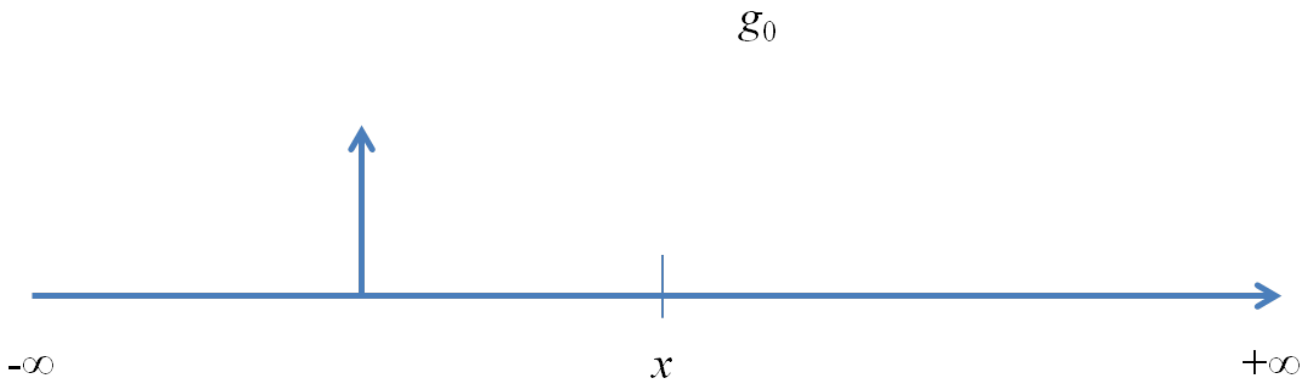



$$\mathcal{E}[P(X < x)] = g_0(-\infty, x] = 0$$

# Near-ignorance solution: Imprecise Dirichlet Process

Keep $s$ fixed and let $g_0$ vary in the set of all probability measures:

$$IDP: \quad \{Dp(s, g_0), \quad g_0 \in \mathbb{P}\}$$

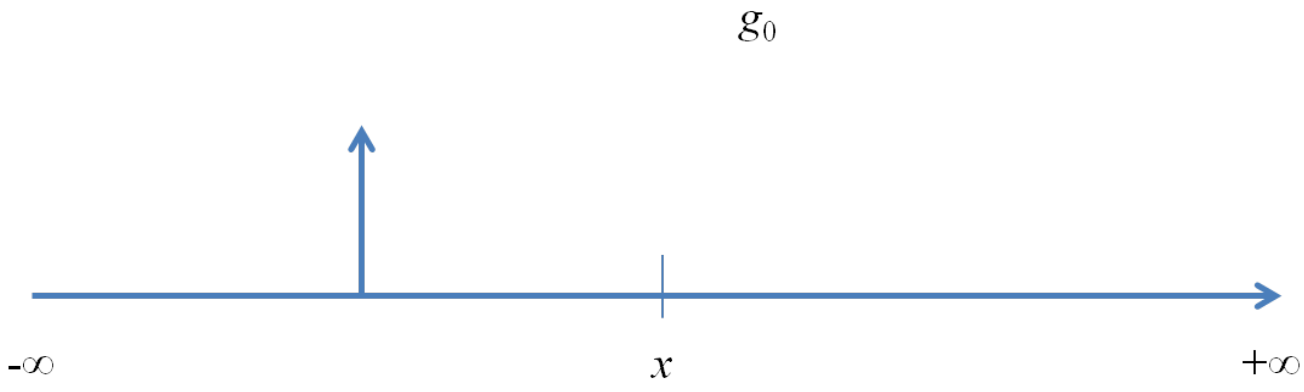



$$\mathcal{E}[P(X < x)] = g_0(-\infty, x] = 0 \qquad \mathcal{E}[P(X < x)] = g_0(-\infty, x] = 1$$

# Near-ignorance solution: Imprecise Dirichlet Process

Keep $s$ fixed and let $g_0$ vary in the set of all probability measures:

$$IDP: \quad \{Dp(s, g_0), \quad g_0 \in \mathbb{P}\}$$




$$\mathcal{E}[P(X < x)] = g_0(-\infty, x] = 0 \qquad \mathcal{E}[P(X < x)] = g_0(-\infty, x] = 1$$

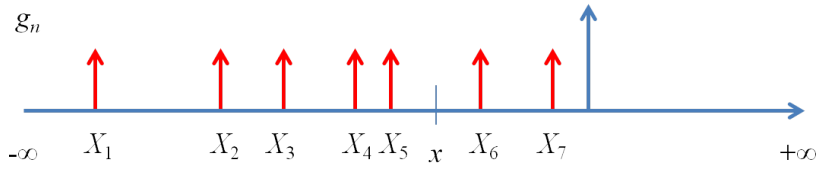
$$0 \leq \mathcal{E}[P(X < x)] \leq 1$$

**No prior information about $P(X < x)$**

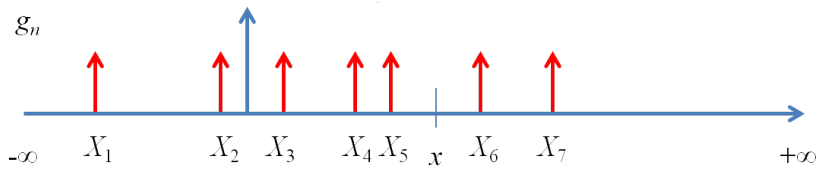
# Imprecise Dirichlet Process - Learning

A posteriori:

LOWER

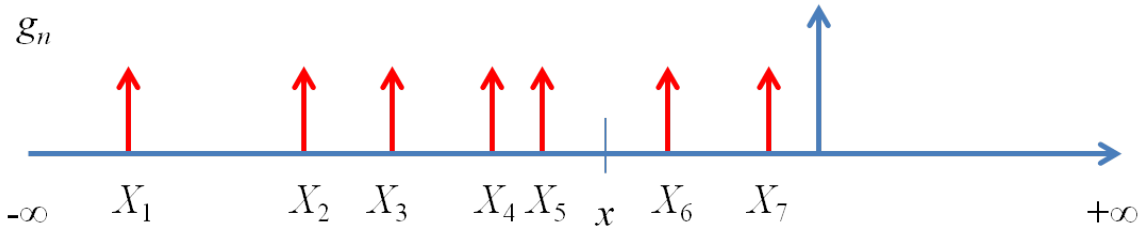



UPPER




$$\frac{n_I}{s+n} < \mathcal{E}[P(X < x)] < \frac{s+n_I}{s+n}$$

# Imprecise Dirichlet Process - Sampling

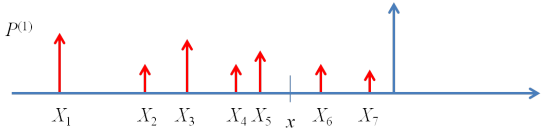


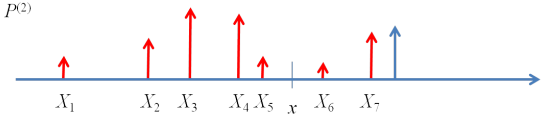

$$g_0 = \delta_{X_0} \Rightarrow g_n \text{ is discrete}$$

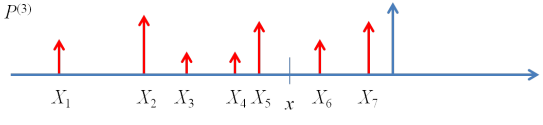
Samples $P^{(k)}$ from $Dp(s+n, g_n)$ have the form

$$P^{(k)} = w_0 \delta_{X_0} + \sum_{i=1}^{n} w_i \delta_{X_i}$$

$$\text{with } (w_0, w_1, \ldots, w_n) \sim \text{Dir}(s, \overbrace{1, \ldots, 1}^{n})$$



# Sign test

$X^n = \{X_1, \ldots, X_n\}$: sequence of observations

$n_l$: # obervations $X_i < 0$ $\quad$ $n_g$: # obervations $X_i \geq 0$

$$H_0: \ P(X < 0) \leq 0.5 \quad H_1: \ P(X < 0) > 0.5$$

Lower

$P(X < 0) \sim \text{Beta}(n_l, s + n_g)$

$\underline{\text{Prob}}(H_1) = \int_{0.5}^{1} \text{Beta}(n_l, s + n_g)$

Upper

$P(X < 0) \sim \text{Beta}(s + n_l, n_g)$

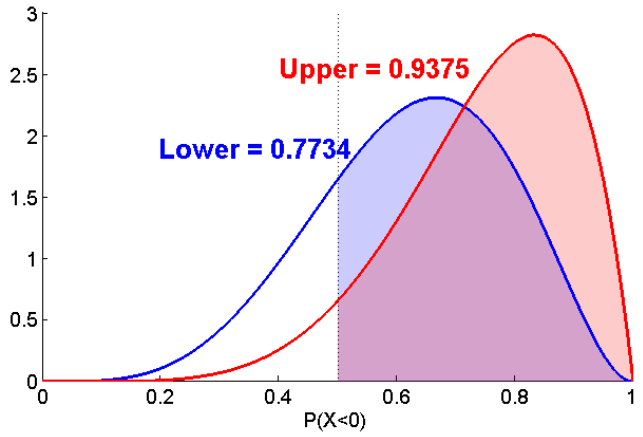
$\overline{\text{Prob}}(H_1) = \int_{0.5}^{1} \text{Beta}(s + n_l, n_g)$

Example:

$n_l = 5,$

$n_g = 2,$

$s = 1$



# The sign test - Decision making

$L_0$: loss associated to action $a_1$ ($=$ accepting $H_1$) if $H_0$ is true (Type I error)

$L_1$: loss associated to action $a_0$ ($=$ accepting $H_0$) if $H_1$ is true (Type II error)

Based on $L_0,\ L_1$ one chooses $a_1$ (accept $H_1$) if

$$\begin{aligned} &\text{Expected loss}|a_1 < \text{Expected loss}|a_0 \\ &\Rightarrow L_0 \text{Prob}(H_0) < L_1 \text{Prob}(H_1) \\ &\Rightarrow \text{Prob}(H_1) > \frac{L_0}{L_0 + L_1} = 1 - \alpha. \end{aligned}$$

What if $\overline{\text{Prob}}(H_1) > 1 - \alpha$ but $\underline{\text{Prob}}(H_1) < 1 - \alpha$?

**The decision is prior-dependent. No robust decision can be taken.**

# Advantages

**Computational tractability:** Sampling from the upper and lower posterior distribution is easier than using stick breaking or other sampling strategies specific to DP.

**Robustness:** When the IDP test is indeterminate the Wilcoxon test virtually behaves as a random guesser (50% of the times issues $H_0$ and the other 50% $H_1$).
The instances that are prior-dependent are somehow critical.
It makes sense to suspend the decisions in those instances.

**Sensitivity analysis:** The maximum value of $s$ that gives a determinate decision can be interpreted as a measure of robustness of the decision.
Even collecting $s$ more observations I will not contradict that decision.

# Wilcoxon rank-sum test

$$\left.\begin{aligned} X^{n_1} &= \{X_1, \ldots, X_{n_1}\} \sim P_X \\ Y^{n_2} &= \{Y_1, \ldots, Y_{n_2}\} \sim P_Y \end{aligned}\right\} \text{sequences of observations from two populations}$$

The aim is to test the equality of distributions for $X$ and $Y$.

$$H_0 : \ P_X = P_Y \quad H_1 : P_X \neq P_Y$$

It uses the linear rank statistic

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{\{X_i < Y_j\}}.$$

The distribution $U$ under the null hypothesis can be computed by considering all the possible random arrangements of the observations in $X^{n_1}$ and $Y^{n_2}$.

# Bayesian rank-sum test

Note that $\frac{U}{n_1 n_2}$ *is the empirical estimate of* $P(X < Y)$

We reformulate the hypothesis

$$H_0 : \ P(X < Y) \leq 0.5 \quad H_1 : P(X < Y) > 0.5$$

Assuming that $P_X \sim Dp(s, g_{0x})$ and $P_Y \sim Dp(s, g_{0y})$, we can the use the DP to make inferences about

$$P(X < Y) = \int P_X(X < y) dP_Y(y)$$

as for instance

$$\mathcal{E}[P(X < Y)|X^{n_1}, Y^{n_2}]$$

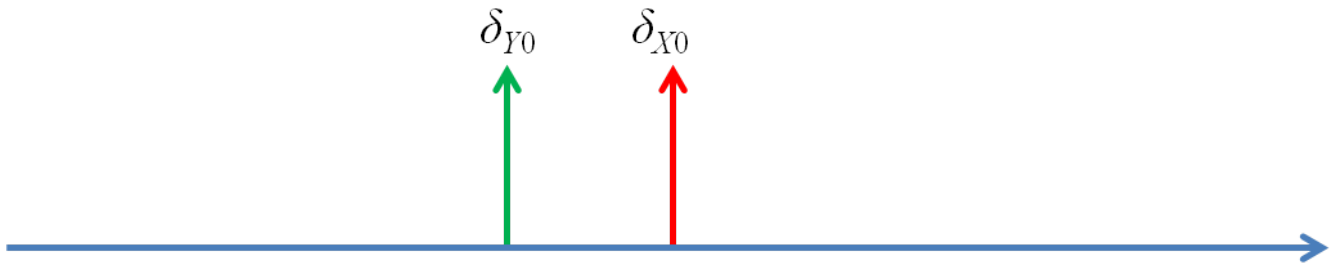
and

$$\text{Prob}\left[P(X < Y) > 0.5|X^{n_1}, Y^{n_2}\right] = \text{Prob}(H_1).$$

# Imprecise rank-sum test - prior

$$P_X \sim Dp(s, g_{0x}) \qquad P_Y \sim Dp(s, g_{0y}),$$

$$g_{0x}, g_{0y} \in \mathbb{P}$$



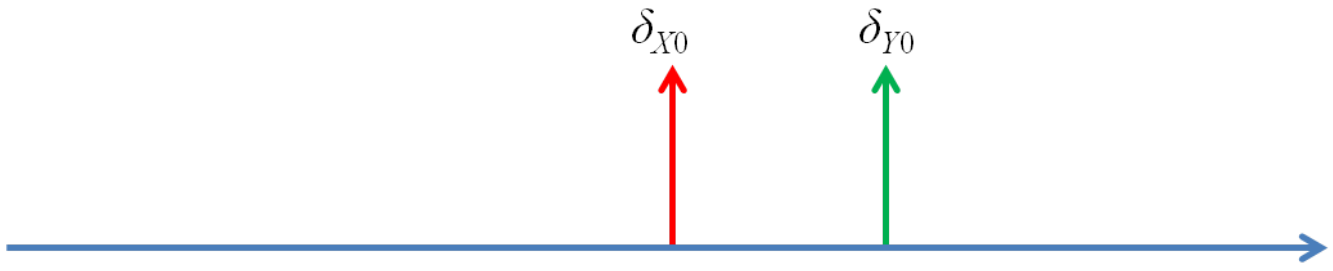

LOWER: $\mathcal{E}[P(X < Y)] = 0$

UPPER: $\mathcal{E}[P(X < Y)] = 1$

# Imprecise rank-sum test - prior

$$P_X \sim Dp(s, g_{0x}) \qquad P_Y \sim Dp(s, g_{0y}),$$

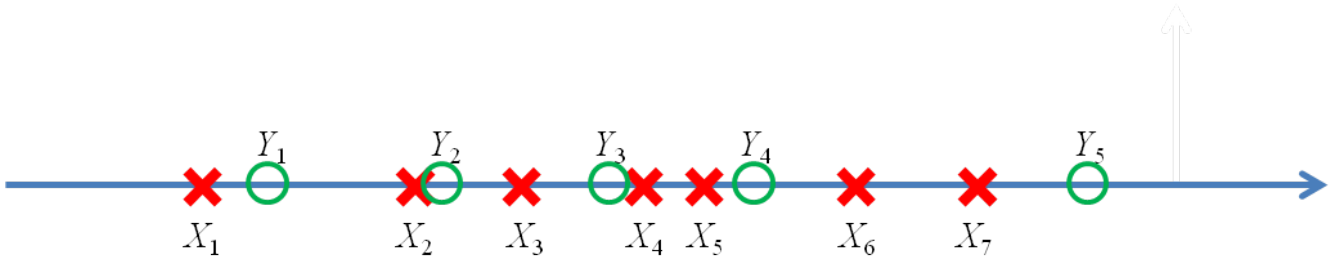
$$g_{0x}, g_{0y} \in \mathbb{P}$$




LOWER: $\mathcal{E}[P(X < Y)] = 0$

UPPER: $\mathcal{E}[P(X < Y)] = 1$

# Imprecise rank-sum test - posterior



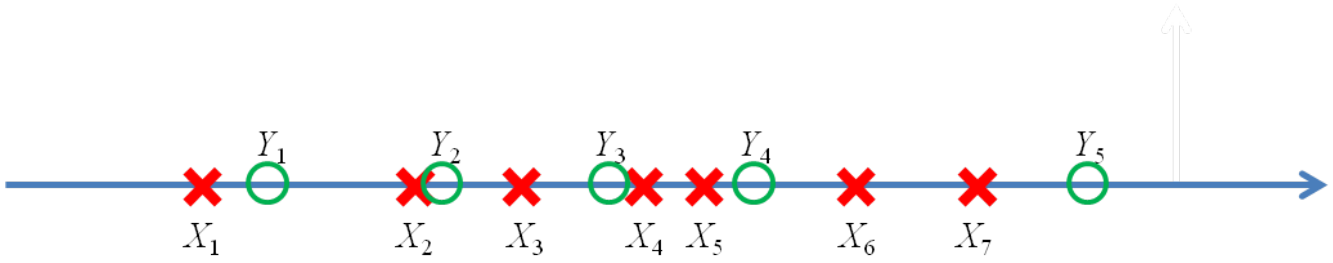

Which prior **minimizes** the probabiliy that $X < Y$? Which prior **maximizes** the probabiliy that $X < Y$?

$$\frac{U}{(s+n_1)(s+n_2)} < \mathcal{E}[P(X<Y)|X^{n_1}, Y^{n_2}] < \frac{U}{(s+n_1)(s+n_2)} + \frac{s(s+n_1+n_2)}{(s+n_1)(s+n_2)}$$

$$P(X<Y) \sim \int P_X^{(k)}(X<y) dP_Y^{(k')}(y)$$

$$= \sum_{i=0}^{n_1} \sum_{i=0}^{n_1} w_{Xi} w_{Yj} I_{\{X_i<Y_j\}}$$

# Imprecise rank-sum test - posterior

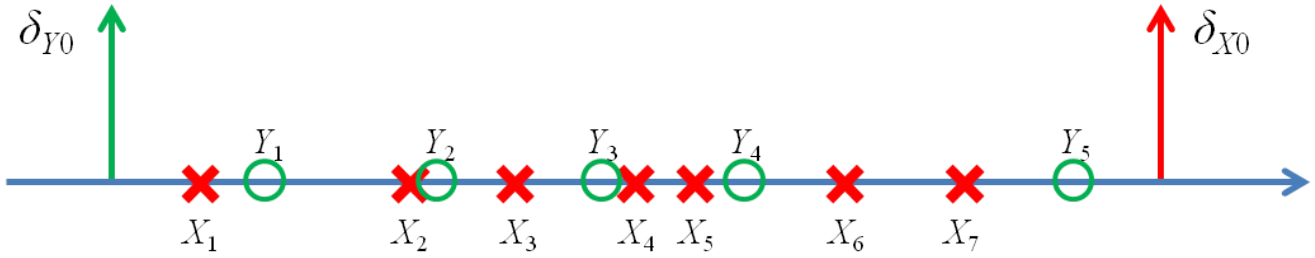



$$\frac{U}{(s+n_1)(s+n_2)} < \mathcal{E}[P(X<Y)|X^{n_1}, Y^{n_2}] < \frac{U}{(s+n_1)(s+n_2)} + \frac{s(s+n_1+n_2)}{(s+n_1)(s+n_2)}$$

$$P(X<Y) \sim \int P_X^{(k)}(X<y) dP_Y^{(k')}(y)$$

$$= \sum_{i=0}^{n_1} \sum_{j=0}^{n_1} w_{Xi} w_{Yj} I_{\{X_i<Y_j\}}$$

# Imprecise rank-sum test - posterior

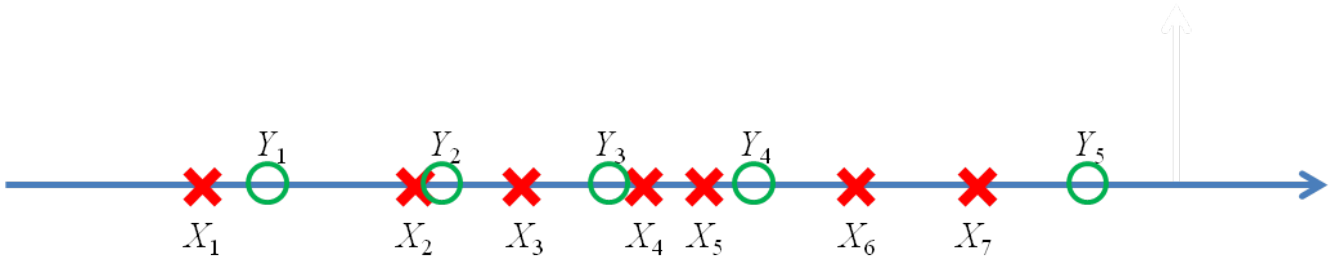



$$\frac{U}{(s+n_1)(s+n_2)} < \mathcal{E}[P(X<Y)|X^{n_1}, Y^{n_2}] < \frac{U}{(s+n_1)(s+n_2)} + \frac{s(s+n_1+n_2)}{(s+n_1)(s+n_2)}$$

$$P(X<Y) \sim \int P_X^{(k)}(X<y) dP_Y^{(k')}(y)$$

$$= \sum_{i=0}^{n_1} \sum_{j=0}^{n_1} w_{Xi} w_{Yj} I_{\{X_i<Y_j\}}$$

# Imprecise rank-sum test - posterior

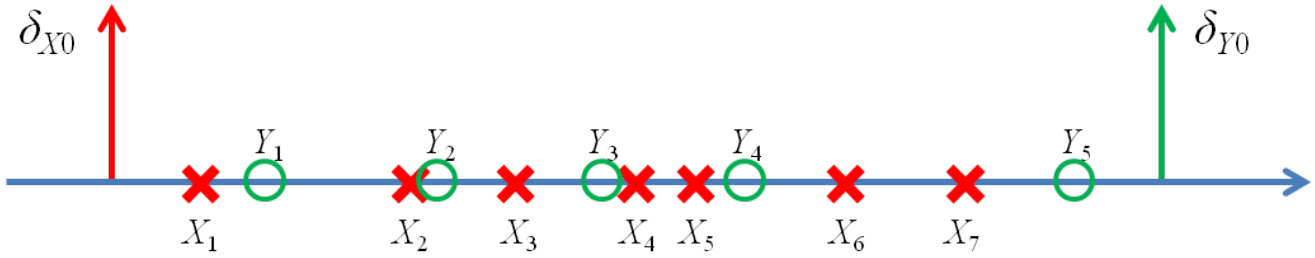



$$\frac{U}{(s+n_1)(s+n_2)} < \mathcal{E}[P(X<Y)|X^{n_1}, Y^{n_2}] < \frac{U}{(s+n_1)(s+n_2)} + \frac{s(s+n_1+n_2)}{(s+n_1)(s+n_2)}$$

$$P(X<Y) \sim \int P_X^{(k)}(X<y) dP_Y^{(k')}(y)$$

$$= \sum_{i=0}^{n_1} \sum_{j=0}^{n_1} w_{Xi} w_{Yj} I_{\{X_i<Y_j\}}$$

# Imprecise rank-sum test - posterior




$$\frac{U}{(s+n_1)(s+n_2)} < \mathcal{E}[P(X<Y)|X^{n_1}, Y^{n_2}] < \frac{U}{(s+n_1)(s+n_2)} + \frac{s(s+n_1+n_2)}{(s+n_1)(s+n_2)}$$

$$P(X<Y) \sim \int P_X^{(k)}(X<y) dP_Y^{(k')}(y)$$

$$= \sum_{i=0}^{n_1} \sum_{j=0}^{n_1} w_{Xi} w_{Yj} I_{\{X_i<Y_j\}}$$

# Advantages

**Computational tractability:** Sampling from the upper and lower posterior distribution is easier than using stick breaking or other sampling strategies specific to DP.

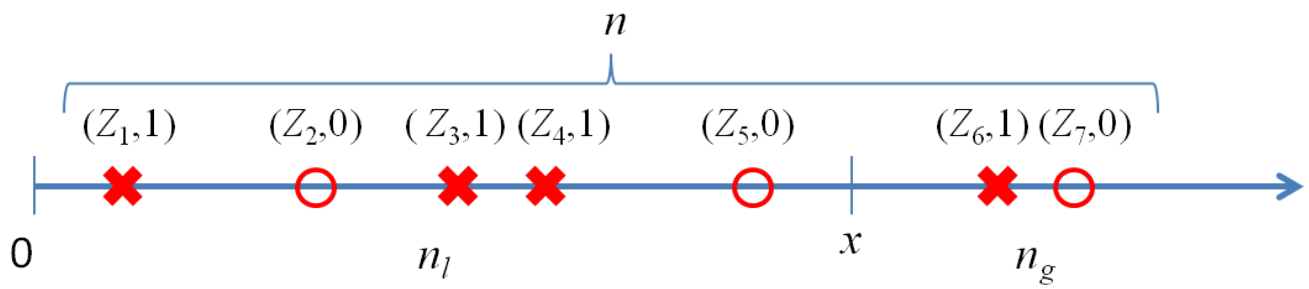
**Robustness:** When the IDP test is indeterminate the Wilcoxon test virtually behaves as a random guesser (50% of the times issues $H_0$ and the other 50% $H_1$).
The instances that are prior-dependent are somehow critical.
It makes sense to suspend the decisions in those instances.

**Sensitivity analysis:** The maximum value of $s$ that gives a determinate decision can be interpreted as a measure of robustness of the decision.
Even collecting $s$ more observations I will not contradict that decision.

**Asymptotic consistency:** If $X$ and $Y$ have same median but different distribution, then the Wilcoxon rank-sum test is not consistent because its null hypothesis is $P_X = P_Y$ but the statistic is an estimator of $P(X < Y) = 0.5$.
The IDP rank-sum test is asymptotically consistent as a test for $P(X < Y)$.

# Survival analysis: right censored data

- $X$ = survival time; $\tilde{X}$ = censoring time
- $Z = \min(X, \tilde{X})$; $d = \delta_{Z=X}$
- $(Z_i, d_i)$ = observation




**POSTERIOR: $P(X > x) \sim$ Mixture of DP (Susarla, Van Ryzin)**

$H(z, d)$ = probability distribution of the bivariate rv $(Z, d)$

PRIOR: $H(z, d) \sim \mathrm{Dp}(s, G(z, d))$

POSTERIOR: $H(z, d) \sim \mathrm{Dp}(s + n, \frac{s}{s+n} G(z, d) + \frac{1}{s+n} \sum \delta_{(Z_i, d_i)})$

**We can sample easily from the posterior distribution of $H(z, d)$** $\underbrace{H^{(k)} \longrightarrow F^{(k)}}$

Peterson, 1977

# Application: numerical simulations

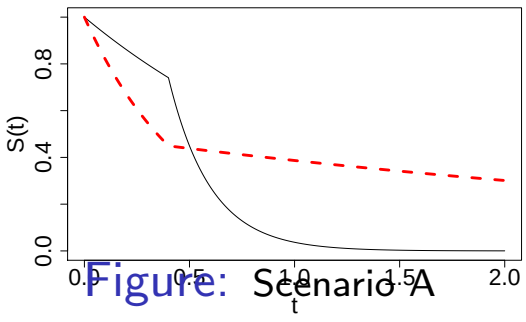



Figure: Scenario A

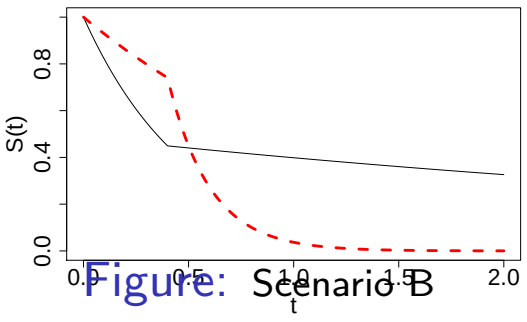



Figure: Scenario B

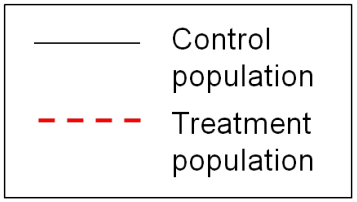



Table: Type-I error ($\alpha = 0.05$)

| | Log-rank | Peto-Peto | IDP ($s = 0.25$) |
|---|---|---|---|
| Scenario A | 0.327 | 0.027 | 0.031 |
| Scenario B | 0.002 | 0.076 | 0.042 |

**Log-rank test:** assumes proportional hazard.

**Generalized Wilcoxon rank-sum test:** weights more early differences.

**IDP test:** Type-I error always $< \alpha$.

# Application: real-world data

- Australian AIDS survival dataset (Ripley and Solomon, 1994).
- We consider five different pairs of groups of individuals and test whether survival is shorter for the first than for the second group.
- According to Ripley and Solomon study:
  - no difference between *Male/Female* and *NSW/VIC* regions,
  - difference between *No-drug/drug* usage, and *Blood/ Haemophilia* transmission.
  - a difference was identified between *QLD/NSW* regions but it was declared *"obscure and requiring further investigation"*.

Table: p-values (Log-rank and generalized Wilcoxon tests) and posterior probabilities (IDP test).

| Group 1 | Group 2 | Log-rank | Peto-Peto | IDP ($s = 0.25$) | $s_{max}$ |
|---|---|---|---|---|---|
| Male | Female | 0.182 (H0) | 0.525 (H0) | 0.579/0.732 (H0) | 5.63 |
| NSW | VIC | 0.228 (H0) | 0.024 (H1) | 0.911/0.920 (H0) | 1.25 |
| No Drug | Drug | 0.011 (H1) | 0.019 (H1) | 0.986/0.990 (H1) | 3.13 |
| Blood | Haemoph. | 0.046 (H1) | 0.007 (H1) | 0.991/0.996 (H1) | 2.08 |
| QLD | NSW | 0.046 (H1) | 0.027 (H1) | 0.927/0.967 ( I ) | - |

# Application: real-world data

- Australian AIDS survival dataset (Ripley and Solomon, 1994).
- We consider five different pairs of groups of individuals and test whether survival is shorter for the first than for the second group.
- According to Ripley and Solomon study:
  - no difference between *Male/Female* and *NSW/VIC* regions,
  - difference between *No-drug/drug* usage, and *Blood/ Haemophilia* transmission.
  - a difference was identified between *QLD/NSW* regions but it was declared *"obscure and requiring further investigation"*.

Table: p-values (Log-rank and generalized Wilcoxon tests) and posterior probabilities (IDP test).

| Group 1 | Group 2 | Log-rank | Peto-Peto | IDP ($s = 0.25$) | $s_{max}$ |
|---|---|---|---|---|---|
| Male | Female | 0.182 (H0) | 0.525 (H0) | 0.579/0.732 (H0) | 5.63 |
| NSW | VIC | 0.228 (H0) | 0.024 (H1) | 0.911/0.920 (H0) | 1.25 |
| No Drug | Drug | 0.011 (H1) | 0.019 (H1) | 0.986/0.990 (H1) | 3.13 |
| Blood | Haemoph. | 0.046 (H1) | 0.007 (H1) | 0.991/0.996 (H1) | 2.08 |
| QLD | NSW | 0.046 (H1) | 0.027 (H1) | 0.927/0.967 ( I ) | - |

# Application: real-world data

- Australian AIDS survival dataset (Ripley and Solomon, 1994).
- We consider five different pairs of groups of individuals and test whether survival is shorter for the first than for the second group.
- According to Ripley and Solomon study:
  - no difference between *Male/Female* and *NSW/VIC* regions,
  - difference between *No-drug/drug* usage, and *Blood/ Haemophilia* transmission.
  - a difference was identified between *QLD/NSW* regions but it was declared *"obscure and requiring further investigation"*.

Table: p-values (Log-rank and generalized Wilcoxon tests) and posterior probabilities (IDP test).

| Group 1 | Group 2 | Log-rank | Peto-Peto | IDP ($s = 0.25$) | $s_{max}$ |
|---|---|---|---|---|---|
| Male | Female | 0.182 (H0) | 0.525 (H0) | 0.579/0.732 (H0) | 5.63 |
| NSW | VIC | 0.228 (H0) | 0.024 (H1) | 0.911/0.920 (H0) | 1.25 |
| No Drug | Drug | 0.011 (H1) | 0.019 (H1) | 0.986/0.990 (H1) | 3.13 |
| Blood | Haemoph. | 0.046 (H1) | 0.007 (H1) | 0.991/0.996 (H1) | 2.08 |
| QLD | NSW | 0.046 (H1) | 0.027 (H1) | 0.927/0.967 ( I ) | - |

# Application: real-world data

- Australian AIDS survival dataset (Ripley and Solomon, 1994).
- We consider five different pairs of groups of individuals and test whether survival is shorter for the first than for the second group.
- According to Ripley and Solomon study:
  - no difference between *Male/Female* and *NSW/VIC* regions,
  - difference between *No-drug/drug* usage, and *Blood/ Haemophilia* transmission.
  - a difference was identified between *QLD/NSW* regions but it was declared *"obscure and requiring further investigation"*.

Table: p-values (Log-rank and generalized Wilcoxon tests) and posterior probabilities (IDP test).

| Group 1 | Group 2 | Log-rank | Peto-Peto | IDP ($s = 0.25$) | $s_{max}$ |
|---|---|---|---|---|---|
| Male | Female | 0.182 (H0) | 0.525 (H0) | 0.579/0.732 (H0) | 5.63 |
| NSW | VIC | 0.228 (H0) | 0.024 (H1) | 0.911/0.920 (H0) | 1.25 |
| No Drug | Drug | 0.011 (H1) | 0.019 (H1) | 0.986/0.990 (H1) | 3.13 |
| Blood | Haemoph. | 0.046 (H1) | 0.007 (H1) | 0.991/0.996 (H1) | 2.08 |
| QLD | NSW | 0.046 (H1) | 0.027 (H1) | 0.927/0.967 ( I ) | - |

# Conclusions

Near-ignorance Dirichlet Process based tests:

- **DP:** the DP allows us to estimate the distribution of the data (no need of assumptions as in the NHST).
- **Decision theory:** the Bayesian approach allows us to formulate the hypothesis test as a decision problem (loss based).
- **Flexibility:** the IDP allows us to start the hypothesis test with very weak prior assumptions (only $s$ must be specified).
- **Robustness:** the IDP has the advantage of producing an indeterminate outcome when the decision is prior-dependent.

# The IDP statistical package


`http://ipg.idsia.ch/software/IDP.php`

The IDP project is still in progress:

- IDP based version of the *Wilcoxon rank-sum test*;
- IDP based version of the *Wilcoxon signed-rank test*;
- IDP based version of the *sign test*;
- IDP for analysis of survival data.

# Learning credal sets from (few) data

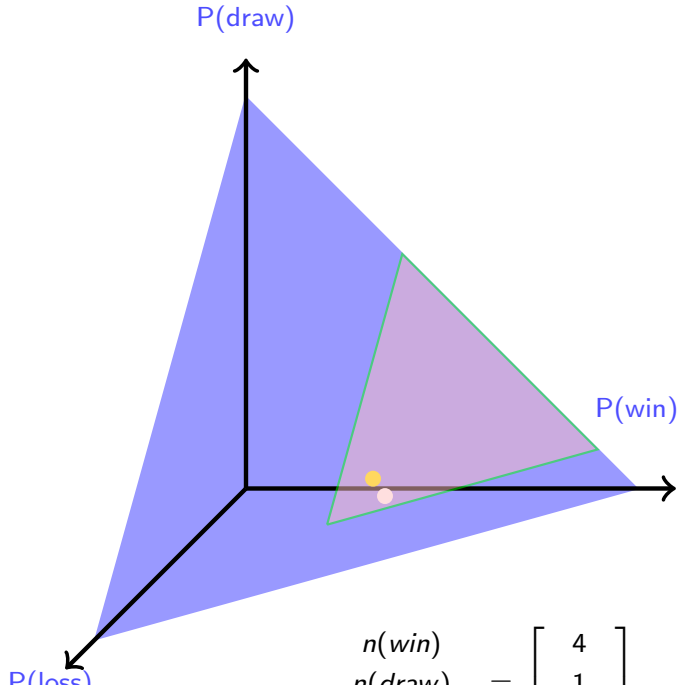
- Learning from data about $X$
- Max lik estimate $P(x) = \frac{n(x)}{N}$
- Bayesian (ESS $s = 2$) $\frac{n(x)+st(x)}{N}$
- Imprecise: set of priors (vacuous $t$)

$$\frac{n(x)}{N+s} \leq P(x) \leq \frac{n(x)+s}{N+s}$$

imprecise Dirichlet model

[Walley & Bernard]

- They a.s.l. and are coherent
- Non-negligible size of intervals only for small $N$
  (Bayesian for $N \to \infty$)




$$\begin{matrix} n(win) \\ n(draw) \\ n(loss) \end{matrix} = \begin{bmatrix} 4 \\ 1 \\ 3 \end{bmatrix}$$

| | | | | |
|---|---|---|---|---|
| 1957: | Spain | vs. | Italy | 5 — 1 |
| 1973: | Italy | vs. | Spain | 3 — 2 |
| 1980: | Spain | vs. | Italy | 1 — 0 |
| 1983: | Spain | vs. | Italy | 1 — 0 |
| 1983: | Italy | vs. | Spain | 2 — 1 |
| 1987: | Spain | vs. | Italy | 1 — 1 |
| 2000: | Spain | vs. | Italy | 1 — 2 |
| 2001: | Italy | vs. | Spain | 1 — 0 |

# Estimation for a categorical variable

- Class $C$, with sample space $\Omega_C = \{c_1, \ldots, c_m\}$.
- $P(c_j) = \theta_j$, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$.
- $n$ i.i.d. observations; $\mathbf{n} = \{n_1, \ldots, n_m\}$.
- *Multinomial* likelihood:

$$L(\boldsymbol{\theta}|\mathbf{n}) \propto \prod_{j=1}^{m} \theta_j^{n_j}$$

- Max. likelihood estimator: $\hat{\theta}_j = \frac{n_j}{n}$.

# Bayesian estimation: Dirichlet prior

- The prior expresses the beliefs about $\boldsymbol{\theta}$, *before* analyzing the data:

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{st_j - 1}.$$

- $s > 0$ is the *equivalent sample size*, which can be regarded as a number of *hidden* instances;
- $t_j$ is the proportion of hidden instances in category $j$.

# Posterior distribution

- Obtained by multiplying likelihood and prior:

$$\pi(\boldsymbol{\theta}|\mathbf{n}) \propto \prod_j \theta_j^{(n_j + st_j - 1)}$$

- *Dir* posteriors are obtained from *Dir* priors (conjugacy).
- Taking expectations:

$$P(c_j|\mathbf{n}, \mathbf{t}) = \frac{n_j + st_j}{n + s}$$

# Example

- $n$=10, $n_1$=4, $n_2$=6, $s$=1.
- The posterior estimate is sensitive on the choice of the prior:

| Prior 1 | Prior 2 |
|---|---|
| $t_1 = 0.5$ | $t_1 = 0.8$ |
| $t_2 = 0.5$ | $t_2 = 0.2$ |
| $\hat{\theta}_1 = \frac{4+0.5}{10+1} = 0.41$ | $\hat{\theta}_1 = \frac{4+0.8}{10+1} = 0.44$ |

Uniform prior is the most common choice.

# Uniform prior *looks* non-informative...

- ... but its estimates depend on the sample space .

| Sample space | $\theta_1, \theta_2$ | $\theta_1, \theta_2, \theta_3$ |
|---|---|---|
| Data | $n_1 = 4$ | $n_1 = 4$ |
| | $n_2 = 6$ | $n_2 = 6$ |
| | - | $n_3 = 0$ |
| Estimate of $\theta_1$ | $\hat{\theta}_1 = \frac{4 + \mathbf{1/2}}{10 + 1} = .41$ | $\hat{\theta}_1 = \frac{4 + \mathbf{1/3}}{10 + 1} = .39$ |

- "Non-informative priors are a lost cause"

L. Wasserman, `normaldeviate.wordpress.com`

# Modelling prior-ignorance: the Imprecise Dirichlet Model

- The IDM is a convex set of Dirichlet prior.
- The set of admissible values for $t_j$ is:

$$\begin{cases} 0 < t_j < 1 \ \ \forall j \\ \sum_j t_j = 1 \end{cases}$$

- This is a model of prior ignorance: a priori nothing is known.

# Learning from data with the IDM

- An *upper* and a *lower* posterior expectation are derived for each chance $\theta_j$.

$$\underline{E}(\theta_j|\mathbf{n}) = \inf_{0<t_j<1} \frac{n_j + st_j}{n+s} = \frac{n_j}{n+s}$$

$$\overline{E}(\theta_j|\mathbf{n}) = \sup_{0<t_j<1} \frac{n_j + st_j}{n+s} = \frac{n_j + s}{n+s}$$

- Upper and lower expectations do *not* depend on the sample space.

# Example

- Binary variable, with $n$=10, $n_1$=4, $n_2$=6, $s$=1.
- The estimates of $\theta_1$ are:

| Bayes $(t_1 = 0.5, t_2 = 0.5)$ | Bayes $(t_1 = 0.8, t_2 = 0.2)$ | IDM |
|---|---|---|
| $\hat{\theta}_1 = \frac{4 + 0.5}{10 + 1}$ | $\hat{\theta}_1 = \frac{4 + 0.8}{10 + 1}$ | $\left[\frac{4}{10 + 1}, \frac{4 + 1}{10 + 1}\right]$ |
| $= .409$ | $= .436$ | $= [.363, .454]$ |

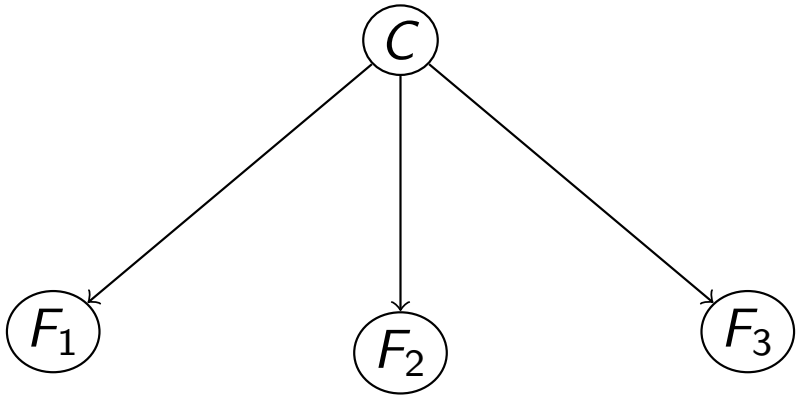
- The interval estimate of the IDM comprises the estimates obtained letting vary each $t_i$ within $(0, 1)$.

# On the IDM

- The estimates are *imprecise*, being characterized by an upper and a lower probability.
- They do not depend on the sample space.
- The gap between upper and lower probability narrows as more data become available.

[Walley, 1996] [Bernard, 2009]

# Naive Bayes (NBC)




- *Naively* assumes the features to be independent given the class.
- NBC is highly biased, but achieves good accuracy, especially on small data sets, thanks to low variance [Friedman,1997].
- Learns from data the joint probability of class and features, decomposed as the marginal probability of the classes and the conditional probability of each feature given the class.

# Joint prior

- $\theta_{c,\mathbf{f}}$ : the unknown joint probability of class and features, which we want to estimate.
- Under naive assumption and Dirichlet prior, the joint prior is:

$$P(\theta_{c,\mathbf{f}}) \propto \prod_{c \in \Omega_C} \theta_c^{st(c)} \prod_{i=1}^{k} \prod_{f \in \Omega_{F_i}} \theta_{(f|c)}^{st(f,c)}.$$

  where $t(f, c)$ is the proportion of hidden instances with $C = c$ and $F_i = f$.
- Let vector $\mathbf{t}$ collect all the parameters $t(c)$ and $t(f, c)$.
- Thus, a joint prior is specified by $\mathbf{t}$.

# Likelihood and posterior

- The likelihood is like the prior, with coefficients $st(\cdot)$ replaced by the $n(\cdot)$.

$$L(\theta|\mathbf{n}) \propto \prod_{c\in\mathcal{C}} \left[ \theta_c^{n(c)} \prod_{i=1}^{k} \prod_{f\in\mathcal{F}_i} \theta_{(f|c)}^{n(f,c)} \right].$$

- The joint posterior $P(\theta_{c,\mathbf{f}}|\mathbf{n},\mathbf{t})$ is like the likelihood, with coefficients $n(\cdot)$ replaced by $st(\cdot) + n(\cdot)$.
- Once $P(\theta_{c,\mathbf{f}}|\mathbf{n},\mathbf{t})$ is available, the classifier is *trained*.

# Issuing a classification

- The value of the features is specified as $\mathbf{f} = (f_i, \ldots, f_k)$.

$$P(c|\mathbf{f}) = P(c) \prod_{i=1}^{k} P(f_i|c)$$

where

$$P(c) = \frac{n(c) + st(c)}{n + s}$$

$$P(f_i|c) = \frac{n(f_i, c) + st(f_i, c)}{n_c + st_c}.$$

- Prior-dependence: the most probable class varies with $\mathbf{t}$.

# Credal classifiers

- Induced using a *set* of priors (*credal set*).
- They separate safe instances from prior-dependent ones.
- On prior-dependent instances: they return a set of classes ( indeterminate classifications ), remaining robust though less informative.

# IDM over the naive topology

We consider a *set* of joint priors, factorized as:

Class :

$$\begin{cases} 0 < t(c) < 1 & \forall c \in \Omega_c \\ \sum_c t(c) = 1 \end{cases}$$

- A priori, $0 < P(c_j) < 1 \ \forall j$.

Features given the class:

$$\begin{cases} \sum_f t(f,c) = t(c) & \forall f, c \\ 0 < t(f,c) < t(c) & \forall f, c \end{cases}$$

- A priori, $0 < P(f|c) < 1, \ \forall c$.

# Naive Credal Classifier (NCC)

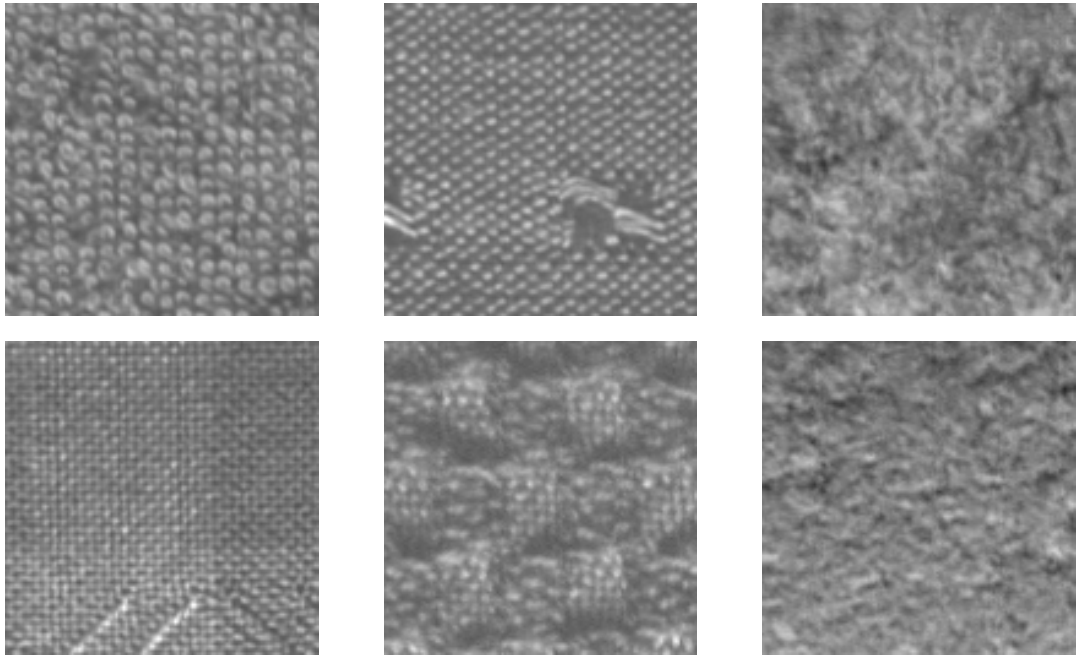
- Uses the IDM to specify a credal set of joint distributions and updates it into a posterior credal set.
- The posterior probability of class $c$ ranges within an *interval*.
- Given feature observation $\mathbf{f}$, class $c'$ credal-dominates $c''$ if for each posterior of the credal set:

$$P(c'|\mathbf{f}) > P(c''|\mathbf{f})$$

# NCC and prior-dependent instances

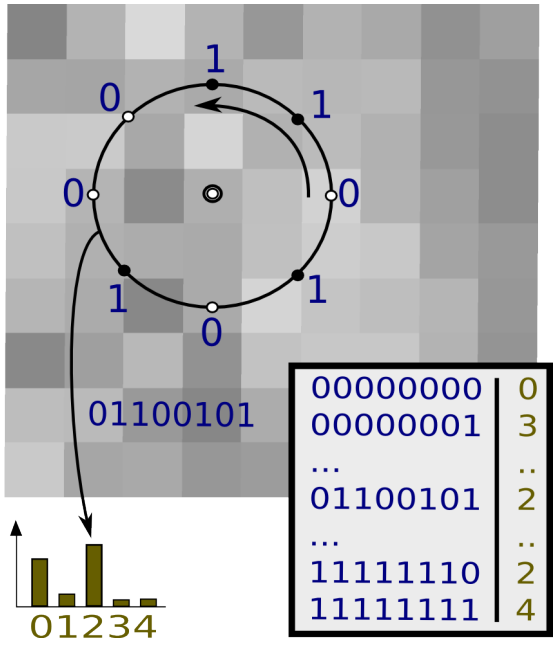
- Credal-dominance is checked by solving an optimization problem.
- NCC eventually returns the *non-dominated* classes:
  - a singleton on the safe instances
  - a set on the prior-dependent ones.
- The next applications shows the gain in reliability due to returning indeterminate classifications on the prior-dependent instances.

# Texture recognition



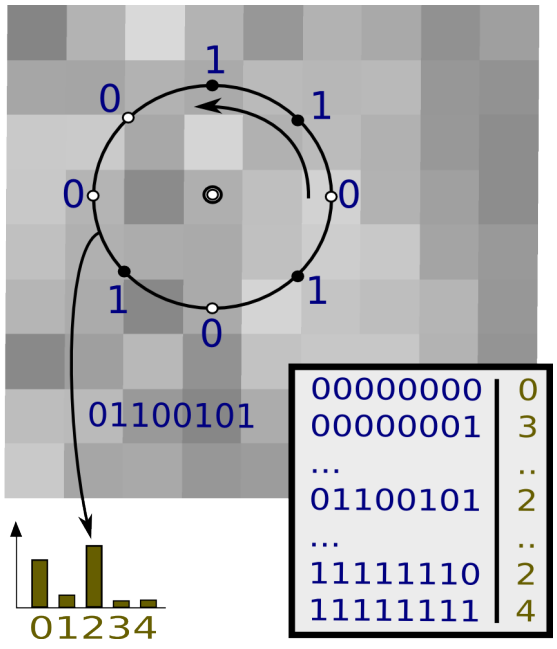
- The OUTEX data sets (Ojala, 2002): 4500 images, 24 classes (textiles, carpets, woods ..).

# Features: Local Binary Patterns (Ojala, 2002)




- The gray level of each pixel is compared with that of its neighbors, resulting in a binary judgment (more intense/ less intense).
- Such judgments are collected in a string for each pixel.

# Local Binary Patterns (2)




- Each string is then assigned to a single category.
- The categories group similar strings: e.g., 00001111 is in the same category of 11110000 for rotational invariance.
- There are 18 categories.
- For each image there are 18 features: the % of pixels assigned to each category.

# Results (Corani et al., BMVC 2010)

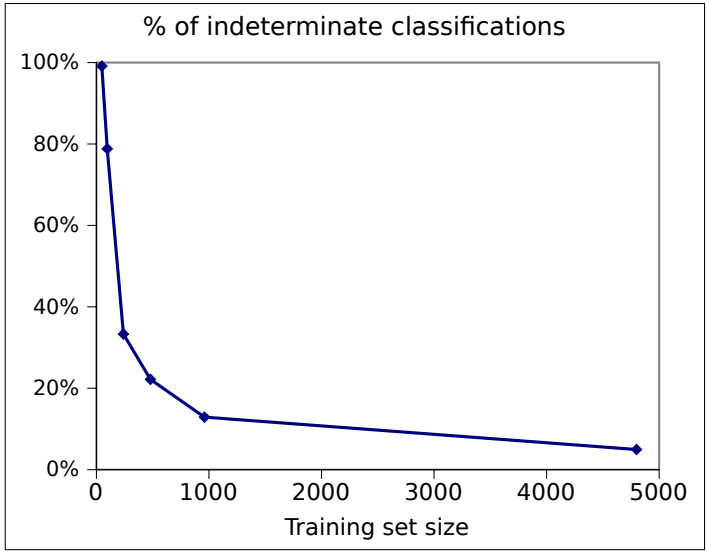
- Accuracy of NBC: 92% (SVMs: 92.5%).
- NBC is highly accurate on the safe instances, but almost random on the prior-dependent ones.

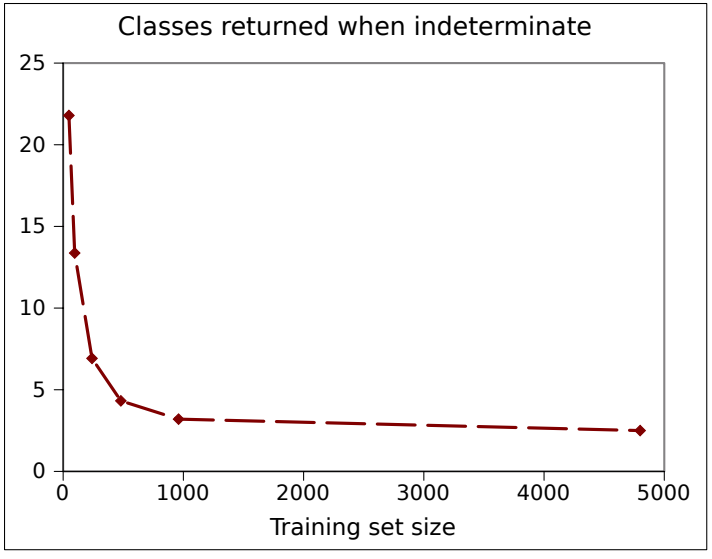
| | *Safe* | *Prior-dependent* |
|---|---|---|
| **Amount%** | 95% | 5% |
| **NBC: accuracy** | 94% | 56% |
| **NCC: accuracy** | 94% | 85% |
| **NCC: non-dom. classes** | 1 | 2.4 |

# Different training set sizes (II)

As $n$ grows there is a *decrease* of both:

- the % of indet. classification;
- the number of classes returned when indeterminate.



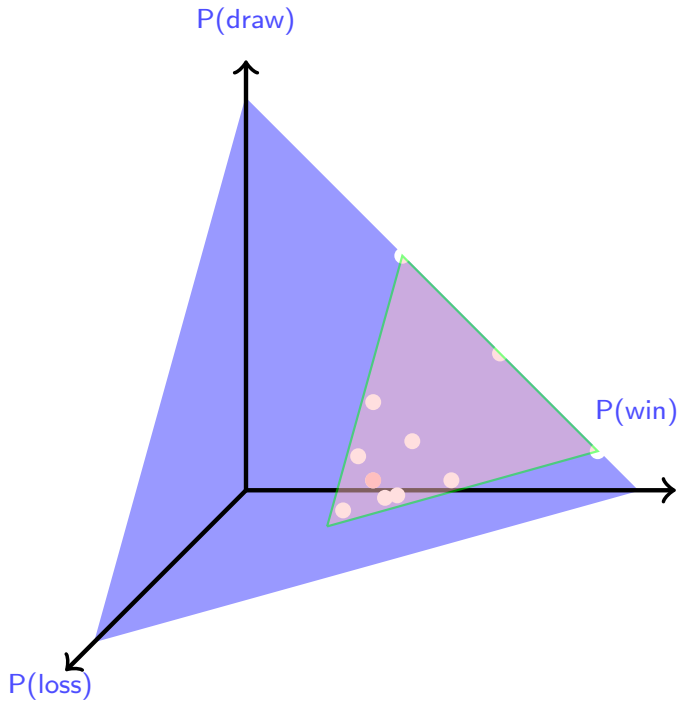



# Naive Bayes + rejection option vs Naive Credal

- Rejection option: reject an instance (no classification) if the probability of the most probable class is below a threshold $p^*$.
- Texture recognition: half of the prior-dependent instances classified by naive Bayes with probability $> 90\%$.
- Rejection rule is not designed to detect prior-dependent instances!

# Learning credal sets from (missing) data

- Coping with missing data?
- Missing at random (MAR) $P(O = *|X = x)$ indep of $X$ Ignore missing data
- Not always the case!
- Conservative updating (de Cooman & Zaffalon) ignorance about the process $P(O|X)$ as a vacuous model
- Consider all the explanations (and take the convex hull)




| | | | | |
|---|---|---|---|---|
| 1957: | Spain | vs. | Italy | 5 — 1 |
| 1973: | Italy | vs. | Spain | 3 — 2 |
| 1980: | Spain | vs. | Italy | 1 — 0 |
| 1983: | Spain | vs. | Italy | 1 — 0 |
| 1983: | Italy | vs. | Spain | 2 — 1 |
| 1987: | Spain | vs. | Italy | 1 — 1 |
| 2000: | Spain | vs. | Italy | 1 — 2 |
| 2001: | Italy | vs. | Spain | 1 — 0 |
| 2003: | Spain | vs. | Italy | $*$ — $*$ |
| 2011: | Italy | vs. | Spain | $*$ — $*$ |

# Ignorance from missing data

- Usually, classifiers *ignore* missing data, assuming them to be MAR (missing at random).
- MAR: the probability of an observation to be missing does not depend on its value or on the value of other missing data.
- The MAR assumption cannot be tested on the incomplete data.

# A non-MAR example: a political poll.

- The right-wing supporters (R) sometimes refuse to answer; left-wing (L) supporters always answer.

| Vote | Answer |
|---|---|
| L | L |
| L | L |
| L | L |
| R | R |
| R | - |
| R | - |

- By ignoring missing data, $P(R) = 1/4$: underestimated!

# Conservative treatment of missing data.

- Consider each possible completion of the data.

| Answer | **D1** | **D2** | **D3** | **D4** |
|---|---|---|---|---|
| L | L | L | L | L |
| L | L | L | L | L |
| L | L | L | L | L |
| R | R | R | R | R |
| - | L | L | R | R |
| - | L | R | L | R |
| $P(R)$ | **1/6** | **1/3** | **1/3** | **1/2** |

- $P(R) \in [1/6,\ 1/2]$; this interval *includes* the real value.
- Application to BNs: Ramoni and Sebastiani (2000).

# Conservative Inference Rule (Zaffalon, 2005)

- MAR missing data are ignored.
- Non-MAR missing data are filled in all possible ways, both in the training and in the test data.
- The replacements exponentially grow with the missing data; yet polynomial time algorithms are available for naive Credal.

[Corani and Zaffalon, 2008, JMLR]

# The conservative treatment of missing data increase indeterminacy.

- Multiple classes are returned if the most probable class depends:
  - on the prior specification *or*
  - on the completion of the non-MAR missing data.
- Declare each feature as MAR or non-MAR depending on domain knowledge, for a good trade-off between robustness and informativeness.

# Accuracy of a credal classifier

*Interval-valued accuracy would be not informative enough (too pessimistic/optimistic)*

- Single-Accuracy (accuracy if a single class is returned)
- Set-Accuracy (accuracy if multiple classes are returned)
- Determinacy (% number of instances with a single class)
- Average output size (average number of classes returned, if indeterminate)
- Bayes-I (accuracy of the Bayesian if credal indeterminate)
- Bayes-D = Single-Accuracy

# Discounted-accuracy

$$\text{d-acc} = \frac{1}{N}\sum_{i=1}^{N}\frac{(accurate)_i}{|Z_i|}$$

- $accurate_i$: whether the set of classes returned on instance $i$ contains the actual class;
- $|Z_i|$: the number of classes returned on the $i$-th instance.
- For a traditional classification, d-acc equals the 0-1 accuracy.

# Doctor random and doctor vacuous.

- Possible diseases:{A,B}.
- Doctor *random*: uniformly random diagnosis.
- Doctor *vacuous*: return both categories.

| Disease | *Random* | | *Vacuous* | |
|---|---|---|---|---|
| | Class. | d-acc | Class. | d-acc |
| A | A | 1 | {A,B} | 0.5 |
| A | B | 0 | {A,B} | 0.5 |
| B | A | 0 | {A,B} | 0.5 |
| B | B | 1 | {A,B} | 0.5 |
| $\overline{\text{d-acc}}$ | | **0.5** | | **0.5** |

# Doctor random vs doctor vacuous: the manager viewpoint.

| Disease | *Random* | | *Vacuous* | |
|---|---|---|---|---|
| | Class. | d-acc | Class. | d-acc |
| A | A | 1 | {A,B} | 0.5 |
| A | B | 0 | {A,B} | 0.5 |
| B | A | 0 | {A,B} | 0.5 |
| B | B | 1 | {A,B} | 0.5 |
| $\overline{\text{d-acc}}$ | | **0.5** | | **0.5** |

- *Assumption*: the hospital profits a quantity of money proportional to the discounted-accuracy.
- After $n$ visits, the profits are:
  - doctor *vacuous*: n/2, with *no variance*.
  - doctor *random*: expected n/2, with *variance* n/4.

# Introducing utility

- Expected utility *increases* with the expected value of the rewards but *decreases* with their variance.
- Any risk-adverse manager prefers doctor vacuous over doctor random.
- And **you** prefer a vacuous over a random diagnosis!
- Idea: quantify such preference through a utility function (Zaffalon et al., IJAR 2012).

# How to design the utility function

| Actual | Predicted | Utility |
|---|---|---|
| A | A | 1 |
| A | B | 0 |

- Utility corresponds with accuracy for traditional classification consisting of a single class.

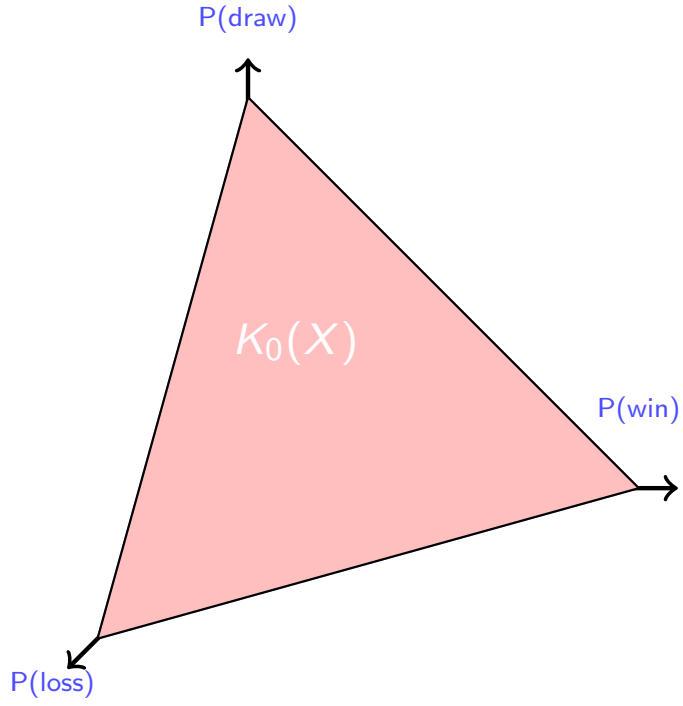
# How to design the utility function

| Actual | Predicted | Utility |
|---|---|---|
| A | B,C | 0 |
| A | A,B | 0.65-0.80 |

- A wrong indeterminate classification has utility 0.
- The utility of an accurate but indeterminate classification consisting of two classes has to be larger than 0.5 ...
- ... otherwise doctor random and doctor vacuous yield the same utility.

# Learning credal sets from experts

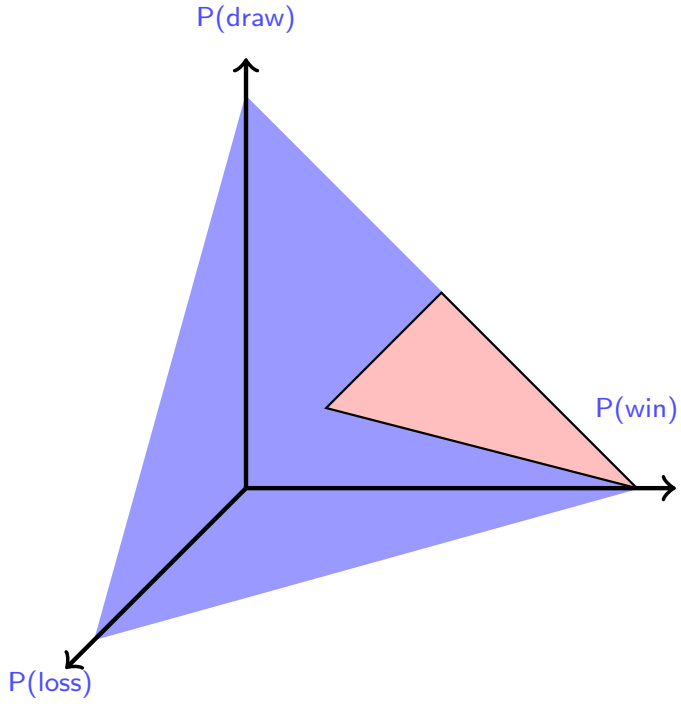
- Modeling ignorance
  - Uniform only models indifference
  - Vacuous credal set
- Expert qualitative knowledge
  - Comparative judgements: win is more probable than draw, which more probable than loss
  - Qualitative judgements: adjective $\equiv$ IP statements




$$K_0(X)=\left\{P(X) \;\middle|\; \begin{array}{l} \sum_x P(x)=1, \\ P(x)\geq 0 \end{array} \right\}$$

# Learning credal sets from experts

- Modeling ignorance
  - Uniform only models indifference
  - Vacuous credal set
- Expert qualitative knowledge
  - Comparative judgements: win is more probable than draw, which more probable than loss
  - Qualitative judgements: adjective $\equiv$ IP statements



# Learning credal sets from experts

- Modeling ignorance
  - Uniform only models indifference
  - Vacuous credal set
- Expert qualitative knowledge
  - Comparative judgements: win is more probable than draw, which more probable than loss
  - Qualitative judgements: adjective $\equiv$ IP statements

**From natural language to linear constraints on probabilities**

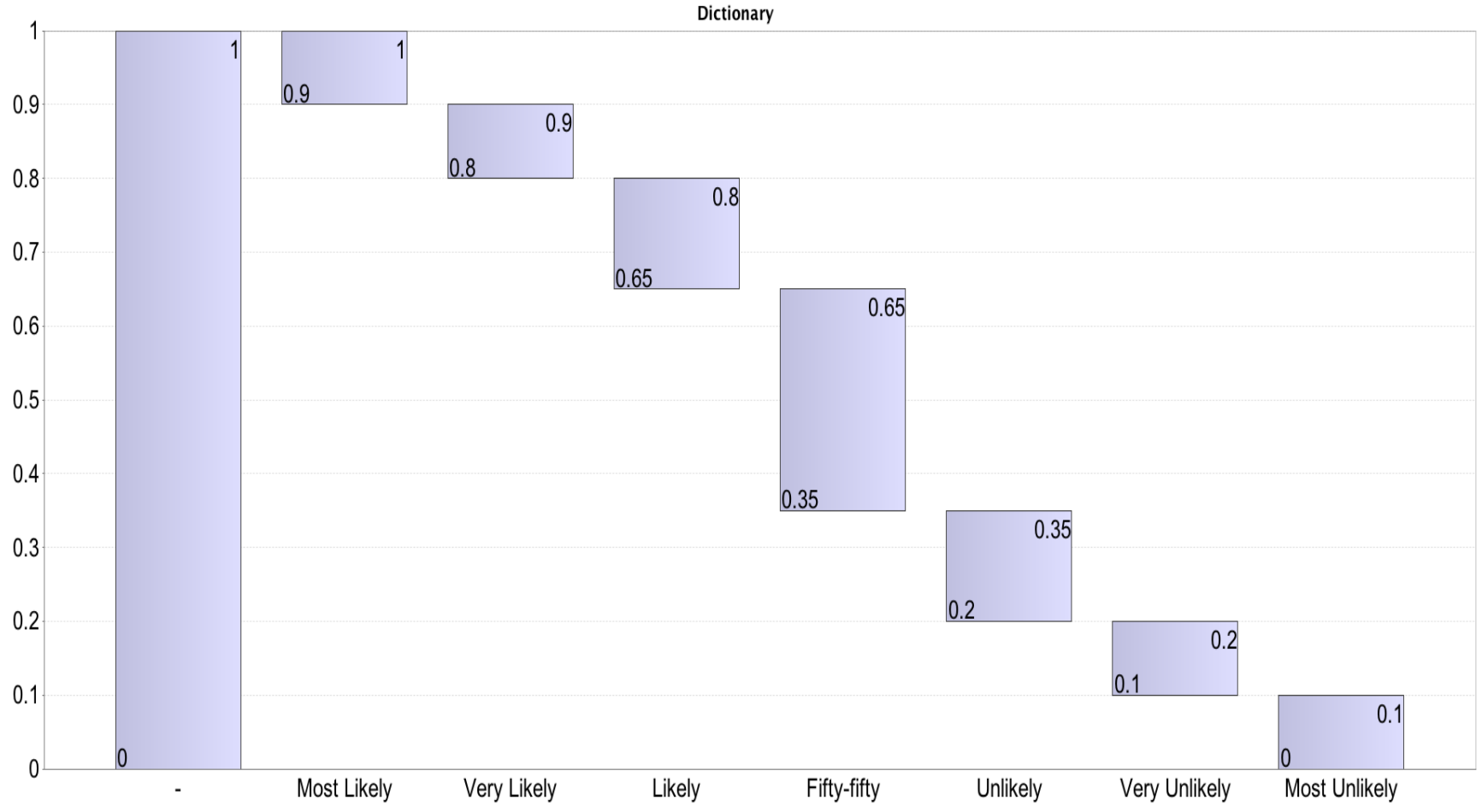
[Walley, 1991]

extremely probable $P(x) \geq 0.98$
very high probability $P(x) \geq 0.9$
highly probable $P(x) \geq 0.85$
very probable $P(x) \geq 0.75$
has a very good chance $P(x) \geq 0.65$
quite probable $P(x) \geq 0.6$
probable $P(x) \geq 0.5$
has a good chance $0.4 \leq P(x) \leq 0.85$
is improbable (unlikely) $P(x) \leq 0.5$
is somewhat unlikely $P(x) \leq 0.4$
is very unlikely $P(x) \leq 0.25$
has little chance $P(x) \leq 0.2$
is highly improbable $P(x) \leq 0.15$
is has very low probability $P(x) \leq 0.1$
is extremely unlikely $P(x) \leq 0.02$

# Credal sets induced by probability intervals

- Assessing lower and upper probabilities: $[l_x, u_x]$, for each $x \in \Omega$
- The consistent credal set is $K(X) := \left\{ P(X) \;\middle|\; \begin{array}{l} l_x \leq P(x) \leq u_x \\ P(x) \geq 0 \\ \sum_x P(x) = 1 \end{array} \right\}$
- Avoiding sure loss implies non-emptiness of the credal set
  $\sum_x l_x \leq 1 \leq \sum_x u_x$
- Coherence implies the reachability (bounds are tight)

$$u_x + \sum_{x' \neq x} l_x \leq 1 \qquad l_x + \sum_{x' \neq x} u_x \geq 1$$


Dictionary
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0
1
0.9
0.9
0.8
0.8
0.65
0.65
0.35
0.35
0.2
0.2
0.1
0.1
0
-
Most Likely
Very Likely
Likely
Fifty-fifty
Unlikely
Very Unlikely
Most Unlikely

# Preventing inconsistent judgements

- E.g., two states of $X$ cannot be both "likely"

  (as this means $P(x) > .65$, $\sum_x P(x) > 1$).
- Reachability constraints

$$\sum_{x \in \Omega_X \setminus \{x'\}} \underline{P}(x) + \overline{P}(x') \leq 1, \quad (1)$$

$$\sum_{x \in \Omega_X \setminus \{x'\}} \overline{P}(x) + \underline{P}(x') \geq 1. \quad (2)$$

- Judgement specification is sequential, the software displays only consistent options