New Perspectives for Sensitivity Analysis

Sebastien Da Veiga – Safran Tech

ETICS 2016

OUTLINE

- **Context**

- **Part I: Screening methods**
  - Model-based approaches
  - Model-free distances

- **Part II: Generalized GSA**
  - Distances between probability distributions
  - RKHS embedding
  - Orthogonal decompositions

CONTEXT

➔ Sensitivity Analysis

- Goal : identify and rank the input parameters according to their impact on the output of a computer code

- Why ?
  - Reduce the output uncertainty efficiently by reducing the uncertainty of the main contributors
  - Improve the knowledge of the physical phenomenon,
  - Simplify the model

- Notations

Computer code

$$Y = \eta(X_1, \ldots, X_d)$$

Output

Input parameters

CONTEXT

➔ Two points of view

- Local Sensitivity: studies the behavior of the output locally around a nominal value of the inputs

$$S_i = \frac{\sigma^2_{X_i}}{\mathrm{Var}(Y)} \left(\frac{\partial \eta(X)}{\partial X_i} |_{X=X_0} \right)^2$$

  - *Easy to compute and apprehend*
  - *But local approach, turns global only if the model is linear*

- Global sensitivity: all input parameters vary in their uncertain domain and we analyze the output variations

There are links between the viewpoints when local sensitivity is repeated (DGSM, Lamboni et al. 2013)

CONTEXT

➔ Global Sensivity Analysis (GSA) – 2 families of methods

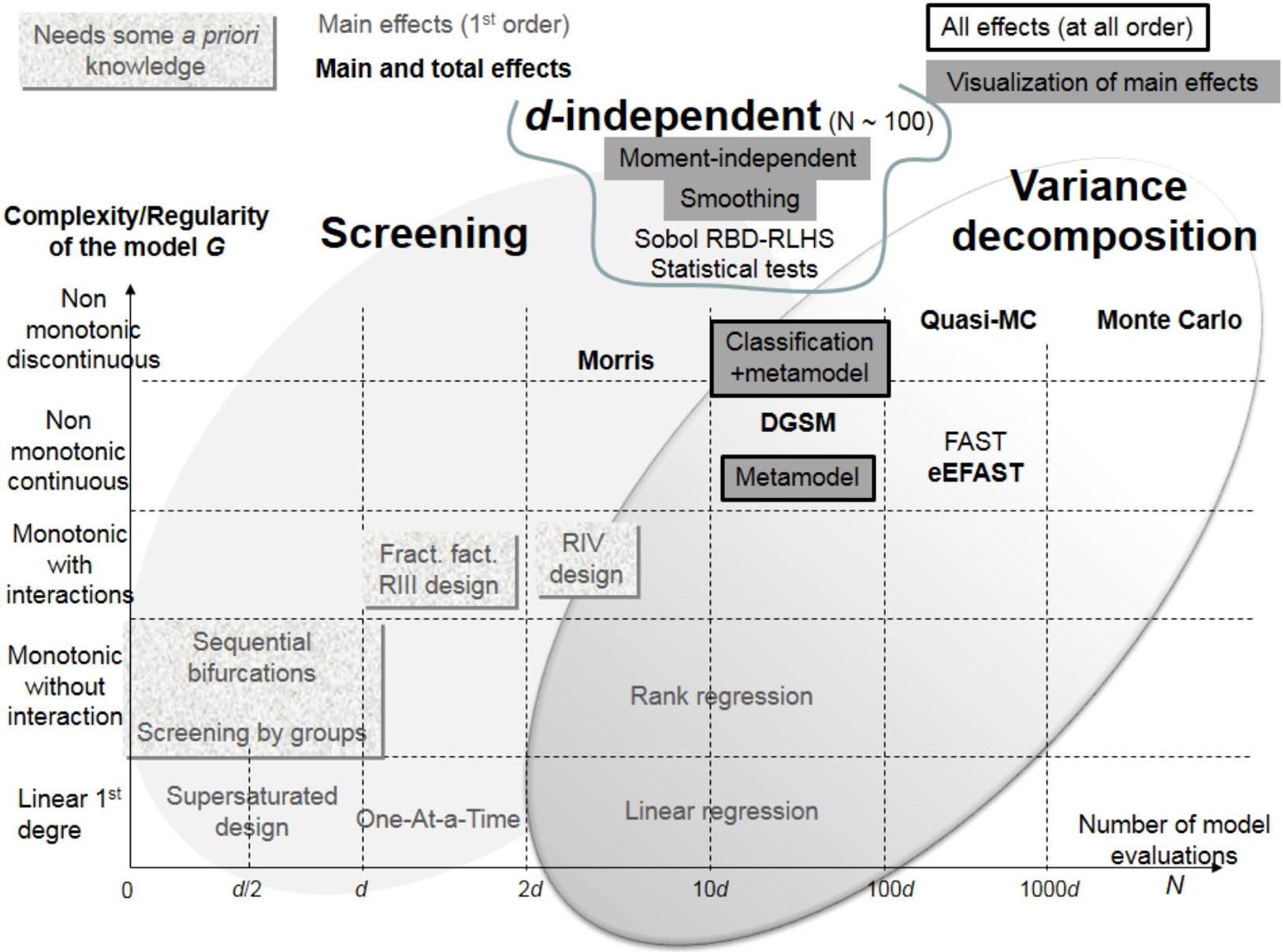
- Screening methods
  - Standard DOEs
  - Sequential bifurcation, …
  - Morris

$$n \approx d/2 - 10d$$

- Quantitative methods based on a variance decomposition
  - Linear regression, SRC, …
  - Sobol indices

$$n \approx 2d - 10^4 d$$

CONTEXT



Iooss & Saltelli 2015

CONTEXT

➔ Goal of this course: alternative methods outside GSA literature

- Screening methods ➔ Feature selection
  - Model-based
  - Model-free

- Quantitative methods based on a variance decomposition ➔ Going beyond the variance
  - Density-based indices
  - Generalized decompositions
  - Link with feature selection

OUTLINE

- Context

- **Part I: Screening methods**
  - Model-based approaches
  - Model-free distances

- Part II: Generalized GSA
  - Distances between probability distributions
  - RKHS embedding
  - Orthogonal decompositions

FEATURE SELECTION

$$Y = \eta(X_1, \ldots, X_d)$$

$$\left(X_1^{(i)}, \ldots, X_d^{(i)}, Y^{(i)}\right)_{i=1,\ldots,n}$$

➔ **Principle: use a sample to identify which inputs are relevant for the output**

- First approach: estimate the relationship between the inputs and the output, **the assumption is that it only involves a small number of inputs (sparsity)**
- Second approach: **rank the inputs according to a relevance criterion** (the criterion is estimated with the sample)

MODEL-BASED APPROACH & SPARSITY

➔ Overview

- Linear hypothesis
  - LASSO
  - With 2nd order interactions
- Nonlinear extensions
  - Sparse additive models
  - With 2nd order interactions
- Non-sparse interlude: random forests

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d$$

➔ **Linear model**

- Solved with ordinary least-squares (with Ridge penalization)

$$\hat{\beta}^{\lambda}_{\mathrm{OLS}} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$\hat{\beta}^{\lambda}_{\mathrm{OLS}} = (X^T X + \lambda I)^{-1} X^T Y$$

- Penalization intensity usually chosen by CV

$$\|u\|_2^2 = \sum_{i=1}^{d} u_i^2$$

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d$$

➔ **Linear model with sparsity**

- Solved with ordinary least-squares penalized by L1-norm (i.e. LASSO)

$$\hat{\beta}^{\lambda}_{\text{LASSO}} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- Several algorithms (LAR, block-coordinate descent,…)
- Penalization intensity usually chosen by CV
- Both LASSO & Ridge: Elastic Net

$$\|u\|_1 = \sum_{i=1}^{d} |u_i|$$

$$\hat{\beta}^{\lambda,\mu}_{\text{EN}} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \mu\|\beta\|_2^2$$

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d$$

➔ **Linear model with sparsity**

- Highly popular
- Efficient solvers
- **But linear assumption only**
- **Only main effects**

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i<j} \theta_{ij} X_i X_j$$

➔ **Linear model with 2nd order interactions & sparsity**

- To the best of my knowledge, a **hierarchy constraint is always imposed**

$$\text{Strong hierarchy: } \theta_{ij} \neq 0 \Rightarrow \beta_i \neq 0 \text{ and } \beta_j \neq 0$$
$$\text{Weak hierarchy: } \theta_{ij} \neq 0 \Rightarrow \beta_i \neq 0 \text{ or } \beta_j \neq 0$$

- **Strong**: there is no interaction between two variables if **at least one** does not have a main effet
- **Weak**: there is no interaction between two variables if **both** don't have a main effect

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i<j} \theta_{ij} X_i X_j$$

➔ **Linear model with 2nd order interactions & sparsity**

- From a general point of view, the problem writes

$$\min_{\beta_0, \beta, \Theta} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda P(\beta, \Theta)$$

- Several choices of penalization function to enforce hierarchy

$$\text{Remark: } \|(u_1, u_2)\|_\gamma + |u_1|, \ \gamma > 1, \ \text{induces } u_2 = 0 \text{ only when } u_1 = 0 \text{ as well}$$

Zhao, Rocha and Yu (2009) – Jenatton, Audibert and Bach (2011)

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i<j} \theta_{ij} X_i X_j$$

➔ **Linear model with 2nd order interactions & sparsity**

- From a general point of view, the problem writes

$$\min_{\beta_0, \beta, \Theta} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda P(\beta, \Theta)$$

- Several choices of penalization function to enforce hierarchy

$$\lambda \sum_{i<j} \left[|\theta_{ij}| + \|(\beta_i, \beta_j, \theta_{ij})\|_{\gamma_{ij}} \right]$$

Zhao, Rocha and Yu (2009) – Composite Absolute Penalties (CAP)

Equivalent formulation in Lim and Hastie (2014) with extension to categorical inputs – glinternet

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i<j} \theta_{ij} X_i X_j$$

➔ **Linear model with 2nd order interactions & sparsity**

- From a general point of view, the problem writes

$$\min_{\beta_0, \beta, \Theta} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda P(\beta, \Theta)$$

- Several choices of penalization function to enforce hierarchy

$$\sum_i \left[\lambda_1 \left\| (\beta_i, \theta_{i:}) \right\|_2 + \lambda_2 \left\| \theta_{i:} \right\|_1 \right]$$

Particular case of Radchenko and James (2010) – VANISH

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i<j} \theta_{ij} X_i X_j$$

➔ **Linear model with 2nd order interactions & sparsity**

- From a general point of view, the problem writes

$$\min_{\beta_0, \beta, \Theta} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda P(\beta, \Theta)$$

- Several choices of penalization function to enforce hierarchy

$$\lambda \sum_{i} \left[\max \left\{ |\beta_i|, \|\theta_{i:}\|_1 \right\} + \|\theta_{i:}\|_1 \right]$$

Bien, Taylor and Tibshirani (2013) – hierNet

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i<j} \theta_{ij} X_i X_j$$

➔ **Linear model with 2nd order interactions & sparsity**

- Focus on hierNet (Bien, Taylor and Tibshirani 2013)

$$\min_{\beta_0, \beta, \Theta} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1$$
$$\text{s.t. } \Theta = \Theta^T$$

All-pairs LASSO, no hierarchy constraint

$$\min_{\beta_0, \beta, \Theta} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1$$
$$\text{s.t. } \Theta = \Theta^T, \ \|\theta_{i:}\|_1 \leq |\beta_i| \text{ for } i = 1, \ldots, d$$

Strong hierarchy constraint added, but not convex

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i<j} \theta_{ij} X_i X_j$$

➔ **Linear model with 2nd order interactions & sparsity**

- Focus on hierNet (Bien, Taylor and Tibshirani 2013)

$$\min_{\beta_0, \beta^+, \beta^- \Theta} \mathcal{L}(\beta_0, \beta^+, \beta^-, \Theta) + \lambda 1^T (\beta^+ + \beta^-) + \frac{\lambda}{2} \|\Theta\|_1$$
$$\text{s.t. } \Theta = \Theta^T, \ \|\theta_{i:}\|_1 \leq \beta_i^+ + \beta_i^-, \ \beta_i^+ \geq 0, \beta_i^- \geq 0 \text{ for } i = 1, \ldots, d$$

Convex relaxation, strong hierarchical LASSO

$$\min_{\beta_0, \beta^+, \beta^- \Theta} \mathcal{L}(\beta_0, \beta^+, \beta^-, \Theta) + \lambda 1^T (\beta^+ + \beta^-) + \frac{\lambda}{2} \|\Theta\|_1$$
$$\text{s.t. } \|\theta_{i:}\|_1 \leq \beta_i^+ + \beta_i^-, \ \beta_i^+ \geq 0, \beta_i^- \geq 0 \text{ for } i = 1, \ldots, d$$

Without symmetry constraint, weak hierarchical LASSO

MODEL-BASED APPROACH & SPARSITY

➔ Linear model with sparsity: practical session

- LASSO & EN: R package glmnet
- Interactions
  - R package hierNet
  - R package glinternet

MODEL-BASED APPROACH & SPARSITY

➔ Overview

- Linear hypothesis
  - LASSO
  - With 2nd order interactions

- Nonlinear extensions
  - Sparse additive models
  - With 2nd order interactions

- Non-sparse interlude: random forests

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \eta_0 + \sum_{i=1}^{d} \eta_i(X_i)$$

➔ **Nonlinear additive model**

- Common in statistics
  - GLM (very specific case), GAM

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \eta_0 + \sum_{i=1}^{d} \eta_i(X_i)$$

➔ **Nonlinear additive model with sparsity**

- Several choices of penalization function to enforce sparsity

$$\min_{\eta_0, \eta_1, \ldots, \eta_d} \mathcal{L}(\eta_0, \eta_1, \ldots, \eta_d) + \lambda \sum_{i=1}^{d} \|\eta_i\|_2$$

Ravikumar et al. (2009)

$$\min_{\eta_0, \eta_1, \ldots, \eta_d} \mathcal{L}(\eta_0, \eta_1, \ldots, \eta_d) + \lambda \sum_{i=1}^{d} \sqrt{\|\eta_i\|_2^2 + \int \eta_i''(x)^2 dx}$$

Meier et al. (2009)

- In practice, expansion of the functions on a basis (splines)

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \eta_0 + \sum_{i=1}^{d} \eta_i(X_i)$$

➔ **Nonlinear additive model with sparsity**

- Particular formulation to distinguish linear and nonlinear effects

$$\eta_i(x) = \beta_i x + \nu_i(x)$$

$$\min_{\eta_0, \beta_i, \nu_i} \mathcal{L}(\eta_0, \beta_i, \nu_i) + \lambda \sum_{i=1}^{d} [\gamma|\beta_i| + (1-\gamma)\|\nu_i\|_2] + \sum_{i=1}^{d} \alpha_i \int \nu_i''(x)^2 dx$$

Chouldechova and Hastie (2015) – GAMSEL

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \eta_0 + \sum_{i=1}^{d} \eta_i(X_i) + \sum_{i<j} \eta_{ij}(X_i, X_j)$$

➔ **Nonlinear additive model with 2nd order interactions & sparsity**

- Several choices of penalization function to enforce sparsity and hierarchy

$$\min_{\eta_0, \eta_1, \ldots, \eta_d} \mathcal{L}(\eta_0, \eta_i, \eta_{ij}) + \lambda \left(\sum_{i=1}^{d} \|\eta_i\|_2 + \sum_{i<j} \|\eta_{ij}\|_2 \right)$$

First proposal in Radchenko and James (2010)
No hierarchy

MODEL-BASED APPROACH & SPARSITY

$$Y = \eta(X_1, \ldots, X_d) = \eta_0 + \sum_{i=1}^{d} \eta_i(X_i) + \sum_{i<j} \eta_{ij}(X_i, X_j)$$

➔ **Nonlinear additive model with 2nd order interactions & sparsity**

- Several choices of penalization function to enforce sparsity and hierarchy

$$\min_{\eta_0, \eta_1, \ldots, \eta_d} \mathcal{L}(\eta_0, \eta_i, \eta_{ij}) + \lambda_1 \sum_{i=1}^{d} \sqrt{\|\eta_i\|_2^2 + \sum_{j \neq i}^{d} \|\eta_{ij}\|_2^2} + \lambda_2 \sum_{i<j} \|\eta_{ij}\|_2$$

Radchenko and James (2010) – VANISH

MODEL-BASED APPROACH & SPARSITY

➔ Nonlinear model with sparsity: practical session

- Main effects only
  - R package SAM (method of Ravikumar et al. 2009)
  - R package gamsel (Chouldechova and Hastie 2015)

- No package available for 2nd order interactions & sparsity, to the best of my knowledge

MODEL-BASED APPROACH & SPARSITY

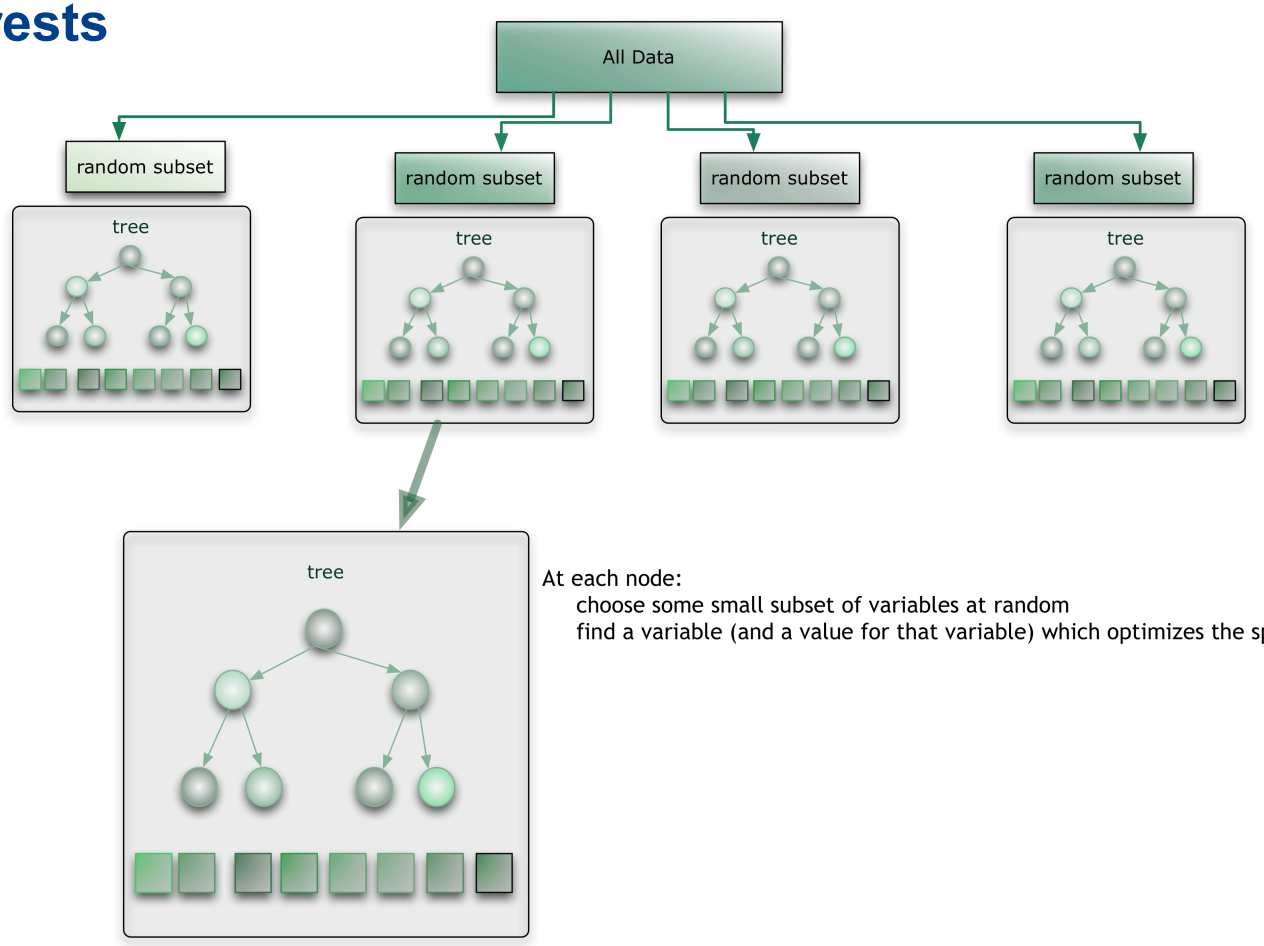
➔ Overview

- Linear hypothesis
  - LASSO
  - With 2nd order interactions
- Nonlinear extensions
  - Sparse additive models
  - With 2nd order interactions
- Non-sparse interlude: random forests

MODEL-BASED APPROACH & SPARSITY

➔ Random forests

-



https://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics/

MODEL-BASED APPROACH & SPARSITY

➔ Random forests

- This mechanism provides a model based on the average over an ensemble of bootstrapped damples
- In parallel, random forests are often used to rank the input variables according to their relevance
- This is done through **variable importance measures**
- Usually, 3 measures are computed
  - The z-score
  - The Gini importance
  - **The permutation importance**

MODEL-BASED APPROACH & SPARSITY

➔ Random forests

- Through bootstrap we can estimate a global prediction error
- Principle: for each tree built, compute the prediction error of this tree on the out-of-bag samples

$$\hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^t)$$

Prediction function of tree « t »

Out-of-bag sample of tree « t »

- The global prediction error is the average over all trees

MODEL-BASED APPROACH & SPARSITY

➔ Random forests

- Permutation importance: if we « break » the link between a variable and the output and we see an increase of the prediction error then the variable is important
  - « Breaking » the link is performed by permuting at random the observations of the variable

- Denote the out-of-bag sample where the observations are permuted $\bar{\mathcal{D}}_n^{tj}$

- The importance measure is then given by

$$I(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{tj}) - \hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{t}) \right]$$

Prediction error on permuted out-of-bag sample

Prediction error on initial out-of-bag sample

MODEL-BASED APPROACH & SPARSITY

➔ Random forests

- Permutation importance: if we « break » the link between a variable and the output and we see an increase of the prediction error then the variable is important
  - « Breaking » the link is performed by permuting at random the observations of the variable

- Denote the out-of-bag sample where the observations are permuted $\bar{\mathcal{D}}_n^{tj}$

- The importance measure is then given by

$$I(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{tj}) - \hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{t}) \right]$$

- **If the input variables are independent, it can be shown that the permutation importance is equal to the unnormalized Sobol index** (up to a factor 2) – Gregorutti (2015)

MODEL-BASED APPROACH & SPARSITY

➔ Random forests: practical session

- R package randomforest

MODEL-BASED APPROACH & SPARSITY

➔ Overview

- Linear hypothesis
  - LASSO
  - With 2nd order interactions
- Nonlinear extensions
  - Sparse additive models
  - **With 2nd order interactions: no R package, any volunteer ? ☺**
- Non-sparse interlude: random forests
- **Other methods not mentioned here: kernel-based**
  - **COSSO (R package cosso), ACOSSO (R code available on Storlie's webpage)**
    - Seems to be limited wrt the number of input variables
  - **Multiple Kernel Learning (MKL) – Hierarchical exploration (Bach 2008, Matlab code available on Bach's webpage)**

OUTLINE

- Context

- **Part I: Screening methods**
  - Model-based approaches
  - Model-free distances

- Part II: Generalized GSA
  - Distances between probability distributions
  - RKHS embedding
  - Orthogonal decompositions

FEATURE SELECTION

$$Y = \eta(X_1, \ldots, X_d)$$

$$\left(X_1^{(i)}, \ldots, X_d^{(i)}, Y^{(i)}\right)_{i=1,\ldots,n}$$

➔ **Principle: use a sample to identify which inputs are relevant for the output**

- First approach: estimate the relationship between the inputs and the output, **the assumption is that it only involves a small number of inputs (sparsity)**

- Second approach: **rank the inputs according to a relevance criterion** (the criterion is estimated with the sample)

MODEL-FREE APPROACH

➔ Generic point of view

- Assume we have a quality criterion which measures the relevance of a subset of input variables

$$\mathcal{Q}(X_u),\ u \subseteq \{1, \ldots, d\}$$

- Feature selection aims at solving

$$u^* = \arg \max_{u \subseteq \{1,\ldots,d\}} \mathcal{Q}(X_u)$$
$$\text{s.t. } |u| \leq s$$

MODEL-FREE APPROACH

$$u^* = \arg \max_{u \subseteq \{1,\dots,d\}} \mathcal{Q}(X_u)$$
$$\text{s.t. } |u| \leq s$$

➔ **Unfortunately this problem is typically NP-hard**

- Impractical in our setting
- It is possible to solve it by investigating all subsets if the number of input variables is small however
- Need to resort to approximate optimization
  - « Marginal »
  - « Greedy »

MODEL-FREE APPROACH

$$u^* = \arg \max_{u \subseteq \{1,\dots,d\}} \mathcal{Q}(X_u)$$
$$\text{s.t. } |u| \leq s$$

➔ **« Marginal » approach**

- Define a threshold (usually depending on the sample size)
- The subset of relevant features is given by

$$u^* = \{i \text{ s.t. } \mathcal{Q}(X_i) \geq t_n\}$$

- But procedures based on marginal computations suffer from the following problems (Fan et al. 2009)
  - Any irrelevant variable that is highly correlated with the set of relevant variables could be selected **(not a problem when inputs are independent)**
  - Marginally uncorrelated variables which are jointly correlated with the output might not be selected **(interactions tend to be undetected)**

MODEL-FREE APPROACH

$$u^* = \arg \max_{u \subseteq \{1,\dots,d\}} \mathcal{Q}(X_u)$$
$$\text{s.t. } |u| \leq s$$

➔ **« Greedy » approach**

- Iterative selection of the relevant variables
- Backward elimination
- Relevant variables are the **last** elements of the output ordered set
- (At each step several variables can be eliminated simultaneously)

Input: The full set of variables $\mathcal{S} = \{1, \dots, d\}$

Output: An ordered set of variables $\mathcal{S}^*$

Init.: $\mathcal{S}^* = \emptyset$

Loop: repeat

$$j^* = \arg\max_j \mathcal{Q}(X_{\mathcal{S}\setminus j})$$

$$\mathcal{S} \leftarrow \mathcal{S}\setminus j^*$$

$$\mathcal{S}^* \leftarrow (\mathcal{S}^*, j^*)$$

MODEL-FREE APPROACH

$$u^* = \arg \max_{u \subseteq \{1,\dots,d\}} \mathcal{Q}(X_u)$$
$$\text{s.t. } |u| \leq s$$

➔ **« Greedy » approach**

- Iterative selection of the relevant variables
- Forward selection
- Relevant variables are the **first** elements of the output ordered set
- (At each step several variables can be selected simultaneously)

Input: The full set of variables $\mathcal{S} = \{1, \dots, d\}$

Output: An ordered set of variables $\mathcal{S}^*$

Init.: $\mathcal{S}^* = \emptyset$

Loop: repeat

$$j^* = \arg\max_j \mathcal{Q}(X_{\mathcal{S}^* \cup j})$$

$$\mathcal{S} \leftarrow \mathcal{S} \backslash j^*$$

$$\mathcal{S}^* \leftarrow (\mathcal{S}^*, j^*)$$

MODEL-FREE APPROACH

$$u^* = \arg \max_{u \subseteq \{1,\dots,d\}} \mathcal{Q}(X_u)$$
$$\text{s.t. } |u| \leq s$$

➔ **Unfortunately this problem is typically NP-hard**

- Marginal or greedy (or combination)
  - Marginal is inexpensive
  - Backward elimination usually provides better features (they are assessed within the context of all others present)
  - Forward selection is computationally more efficient

- **The key point is to choose wisely the quality criterion**

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$\mathcal{Q}(X_u),\ u \subseteq \{1, \ldots, d\}$$

➔ **Historical approach relies on correlations**

- For a subset of only one variable

$$\mathcal{Q}(X_i) = \mathrm{Cor}(X_i, Y)$$

- Generalization for a subset with partial correlations
- **Obviously this approach is limited since it only detects variables which are linearly dependent with the output**

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$\mathcal{Q}(X_u),\ u \subseteq \{1,\dots,d\}$$

➔ **Further advances investigate mutual information**

$$\begin{aligned}\mathcal{Q}(X_u) &= \mathrm{MI}(X_u, Y) \\ &= \int_{\mathbb{R}^{|u|}\times\mathbb{R}} f_{X_u,Y}(x_u,y)\log\frac{f_{X_u,Y}(x_u,y)}{f_{X_u}(x_u)f_Y(y)}dx_u dy\end{aligned}$$

- **Curse of dimensionality (density estimation)**
- Complete feature selection procedure with marginal MI only described in the well-known paper of Peng et al. (2005)

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$\mathcal{Q}(X_u),\ u \subseteq \{1, \ldots, d\}$$

➔ What should we look for ?

- The measure must be able to detect nonlinear effects unlike correlation
  - Just like MI, which is a measure of the dependence between random vectors
- The measure must not rely on density estimation
  - Many dependence measures exist, but almost all of them imply estimating a density

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$\mathcal{Q}(X_u),\ u \subseteq \{1, \ldots, d\}$$

- **Recent proposals are based on measures of dependence between random vectors**
  - Assume we have two random vectors

$$X \in \mathbb{R}^p \text{ and } Y \in \mathbb{R}^q$$

  - Conditions for independence

$$\begin{aligned} X \perp Y &\iff f_{XY}(x,y) = f_X(x) f_Y(y) \\ &\iff \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\ &\iff \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g, h \in \mathcal{F} \end{aligned}$$

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad \boxed{f_{XY}(x,y) = f_X(x) f_Y(y)} \\
&\Longleftrightarrow \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\Longleftrightarrow \quad \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}
\end{aligned}
$$

➔ **Comparing these two densities is what MI does !**

$$
\mathrm{MI}(X,Y) = \int_{\mathbb{R}^p \times \mathbb{R}^q} f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy
$$

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \boxed{\Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s)} \\
&\Longleftrightarrow \quad \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}
\end{aligned}
$$

➔ **How to compare characteristic functions ?**

- For a complex function, define the weighted L2 norm

$$\|\gamma(t,s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(t,s)|^2 w(t,s)\,dt\,ds$$

- A dependence measure may then be

$$\mathcal{V}^2(X,Y;w) = \|\Phi_{XY}(t,s) - \Phi_X(t)\Phi_Y(s)\|_w^2$$

- Central tool introduced by Székely et al. (2007): choose a smart weighting function

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \boxed{\Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s)} \\
&\Longleftrightarrow \quad \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}
\end{aligned}
$$

➔ **Székely et al. (2007)**

- Crucial integration lemma

$$\int_{\mathbb{R}^d} \frac{1-\cos\langle t,x\rangle}{|t|_p^{p+\alpha}} dt = C(d,\alpha)|x|_p^{\alpha} \text{ for } 0<\alpha<2$$

- Choose the weighting function equal to

$$w(t,s) = \left(C(p,1)C(q,1)|t|_p^{p+1}|s|_q^{q+1}\right)^{-1}$$

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\iff \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\iff \quad \boxed{\Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s)} \\
&\iff \quad \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}
\end{aligned}
$$

➔ **Székely et al. (2007)**

- The dependence measure, called Distance Covariance (dCov), is then defined as the square root of

$$\mathcal{V}^2(X,Y) = \frac{1}{C(p,1)C(q,1)} \int_{\mathbb{R}^{p+q}} \frac{|\Phi_{XY}(t,s) - \Phi_X(t)\Phi_Y(s)|^2}{|t|_p^{p+1}|s|_q^{q+1}} dtds$$

- Why is this smart ?
  - It involves characteristic functions which also suffer from the curse of dimensionality, right ?

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \boxed{\Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s)} \\
&\Longleftrightarrow \quad \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}
\end{aligned}
$$

➔ **Székely et al. (2007)**

- Thanks to the previous lemma, we have

$$
\begin{aligned}
\mathcal{V}^2(X,Y) = \mathbb{E}_{XX'YY'}\left[|X-X'|_p|Y-Y'|_q\right] + \mathbb{E}_{XX'}\left[|X-X'|_p\right]\mathbb{E}_{YY'}\left[|Y-Y'|_q\right] \\
-2\mathbb{E}_{XY}\left[\mathbb{E}_X\left[|X-X'|_p\right]\mathbb{E}_Y\left[|Y-Y'|_q\right]\right]
\end{aligned}
$$

- Only expectations !

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \boxed{\Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s)} \\
&\Longleftrightarrow \quad \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}
\end{aligned}
$$

➔ **Székely et al. (2007)**

- The final dependence measure is the distance correlation (dCor), which is defined as the square root of

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}}$$

- Simple estimator given in Székely et al. (2007)
- Generalizations in subsequent papers of Székely & Rizzo (large dimension, time series, …)
- R package energy

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\iff \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\iff \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\iff \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**

- Forget about the dependence measure for now
- « Simpler » problem
  - Assume we have two random vectors

$$U \in \mathbb{R}^d \sim \mathbb{P}, \; V \in \mathbb{R}^d \sim \mathbb{Q}$$

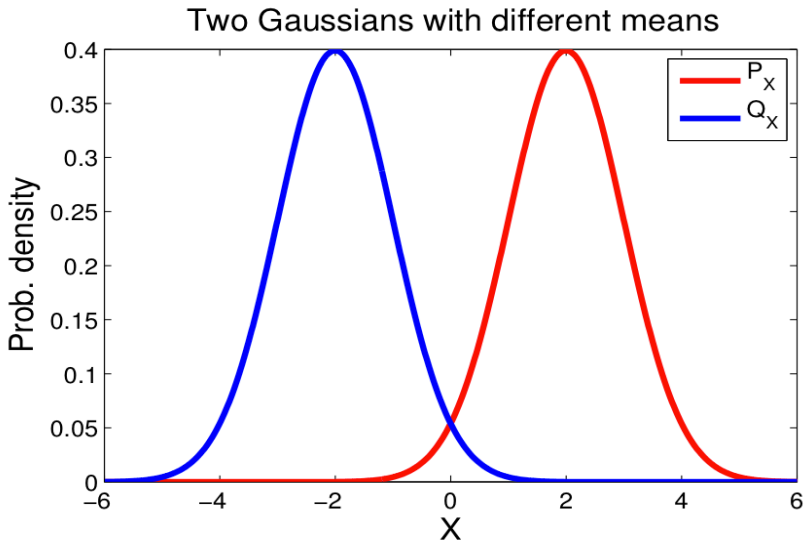
  - We want to test if their probability distributions are the same

$$\mathbb{P} = \mathbb{Q} \; ?$$

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\Longleftrightarrow \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**



$$\mathbb{P} = \mathbb{Q} \ ?$$

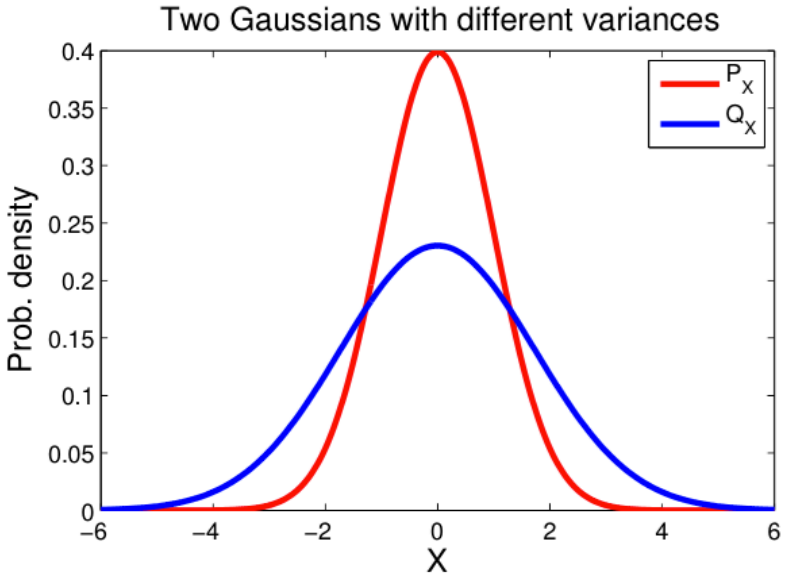
EASY

Gretton 2012

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\Longleftrightarrow \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**



$$\mathbb{P} = \mathbb{Q} \ ?$$

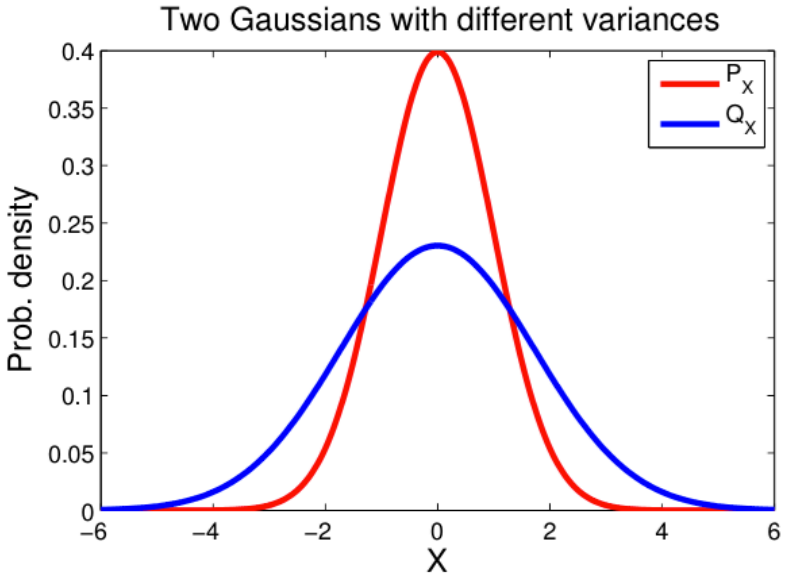
NOT EASY

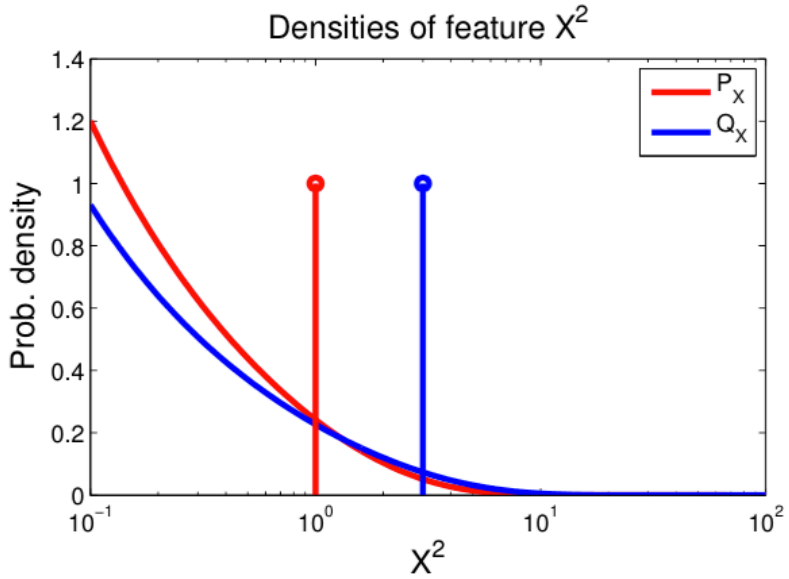
Gretton 2012

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\Longleftrightarrow \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**





$$\mathbb{P} = \mathbb{Q} \ ?$$

EASY

Gretton 2012

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\iff \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\iff \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\iff \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**

- The idea is to compare the probability measures through their means taken on several transformations

- More on this in Part II, but some hints for the moment
  - No need to specify manually all these transformations
  - Only a kernel must be chosen, as long as it is characteristic
    - E.g. Gaussian, Laplace
  - The hyperparameter(s) of the kernel must be provided (rules of thumb)

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\Longleftrightarrow \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**

- To compare two probability measures, the metric is defined as

$$
\begin{aligned}
\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{XX'}[k(X,X')] + \mathbb{E}_{ZZ'}[k(Z,Z')] \\
-2\mathbb{E}_{XZ}[k(X,Z)]
\end{aligned}
$$

Smola et al. 2007

- Estimation through Gram matrices X and Z only

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\Longleftrightarrow \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\Longleftrightarrow \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\Longleftrightarrow \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**

- Finally to measure the dependence between two random vectors, just apply this measure to compare the two probability measures

$$\mathbb{P}_{XY} \text{ and } \mathbb{P}_X \times \mathbb{P}_Y$$

- This gives rise to the so-called Hilbert-Schmidt Independence Criterion (HSIC)

$$\mathrm{HSIC}(X,Y) = \mathrm{MMD}^2(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y)$$

Gretton 2005

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$
\begin{aligned}
X \perp Y \quad &\iff \quad f_{XY}(x,y) = f_X(x) f_Y(y) \\
&\iff \quad \Phi_{XY}(t,s) = \Phi_X(t)\Phi_Y(s) \\
&\iff \quad \boxed{\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \text{ for } g,h \in \mathcal{F}}
\end{aligned}
$$

➔ **A measure relying on moments directly ?**

- Once kernels are defined for inputs and outputs, HSIC only involves moments

$$
\begin{aligned}
\mathrm{HSIC}(X,Y) = \mathbb{E}_{XX'YY'}[k(X,X')l(Y,Y')] + \mathbb{E}_{XX'}[k(X,X')]\mathbb{E}_{YY'}[l(Y,Y')] \\
-2\mathbb{E}_{XY}[\mathbb{E}_X[k(X,X')]\mathbb{E}_Y[l(Y,Y')]]
\end{aligned}
$$

- This is actually a generalization of dCov (Sejdinovic et al. 2013)
  - Choose specific kernels and you will recover dCov

MODEL-FREE APPROACH: CHOICE OF RELEVANCE MEASURE

$$\mathcal{Q}(X_u),\ u \subseteq \{1, \ldots, d\}$$

➔ What we have seen

- (Partial) Correlation
- Mutual information
- **dCov**
- **HSIC**

- **The last two have been used recently in several feature selection problems**

➔ Not mentioned here

- A mixture of HSIC & Lasso by Yamada et al. (2014)
- Matlab code available at Yamada's webpage

MODEL-FREE APPROACH

➔ Practical session

- Implementation of marginal, backward & forward approaches
- With dCov and HSIC
- (Multiple eliminations/selections at each step)