varrank:

an R package for variable ranking based on mutual information with applications to observed systems epidemiology

Gilles Kratzer[1], Reinhard Furrer[1,2]

[1]Department of Mathematics, [2]Department of Computational Science; University of Zurich (Switzerland)

Contact: gilles.kratzer@math.uzh.ch

Motivation

- In system epidemiology, the typical set of possible variables is large
- Classical approaches for variables selection:[1]
  - Prior scientific knowledge: 29%
  - Change of estimate: 18%
  - Stepwise model selection: 16%

No prior model?

Not one outcome experiment?

Summary

- Variable ranking based on a **set of variable of importance**
- **Model free**
- Flexible implementation of the **mRMRe[2]** algorithm
- Mixture of variables (**continuous** and **discrete**)
- Discretisation through rule/clustering
- **varrank** is distributed as an R package

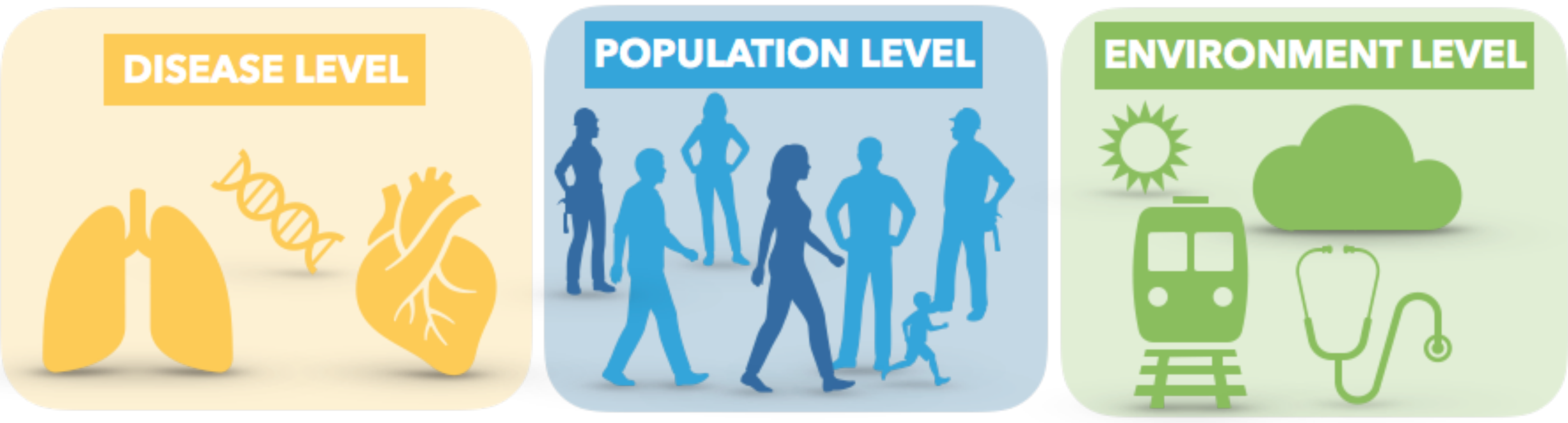
https://CRAN.R-project.org/package=varrank

Results

- Can be used as a ranker/selector
- **Multiple** variable of importance possible
- Highly parametrizable
- Nice graphical output

Many synergies with network modelling approaches applied to systems epidemiology

Why systemic thinking?



- Systems epidemiology implies contributions at different levels
- Confounding factors
- Complex dependance structure
- Multicollinearity

How to perform variable ranking using varrank?

$$\mathrm{H}(X) = \sum_{n=1}^{N} P(x_n) \log P(x_n)$$

Average amount of information of one RV

$$\mathrm{MI}(X;Y) = \sum_{n=1}^{N}\sum_{m=1}^{M} P(x_n; y_m) \log \frac{P(x_n; y_m)}{P(x_n)P(y_m)}$$

Mutual dependence between two RV

Difference (mid) or quotient (miq)

Greedy search

Forward - argmax

$$\mathrm{score}_i = \underbrace{\mathrm{MI}(f_i; \mathbf{C})}_{\text{Relevance}} - \beta \sum_{\mathbf{S}} \underbrace{\alpha(f_i, f_s, \mathbf{C})}_{\text{Normalization}} \underbrace{\mathrm{MI}(f_i; f_s)}_{\text{Redundancy}}$$

Backward - argmax

Discretization

Estévez and al. (2009)

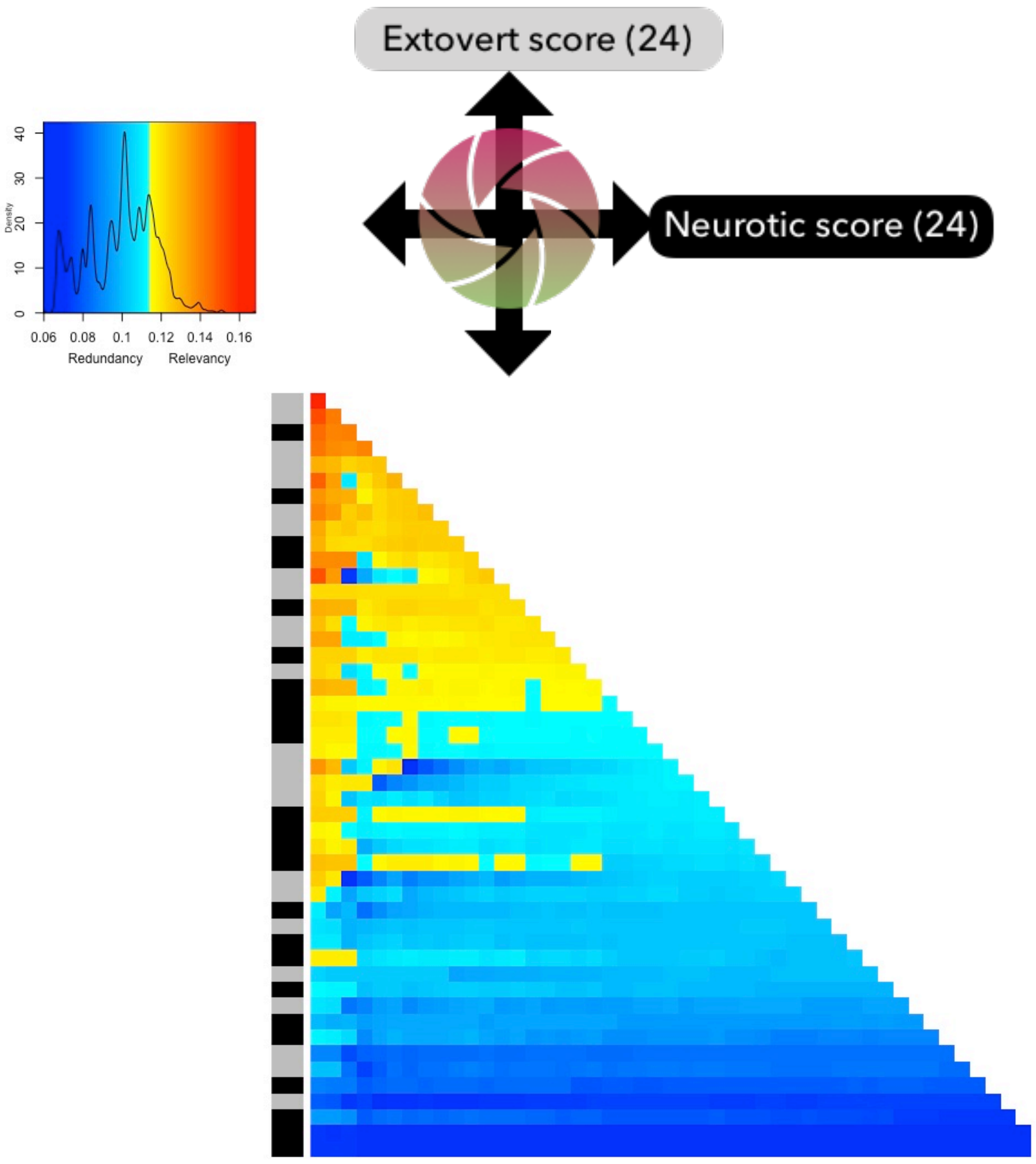
$$\beta = 1/|\mathbf{S}| \text{ and } \alpha(f_i, f_s, \mathbf{C}) = \frac{1}{\min(\mathrm{H}(f_i), \mathrm{H}(f_s))}$$

Illustrative example

Eysenck Personality Inventory (EPI)[3], is and has been a very frequently administered personality test with 57 measuring two broad emotional dimensions

57 variables, n = 3570 observations (no missing data)

- Extraversion-Introversion
- Stability-Neuroticism
- Lie scale (variable of importance)



References

1. S. Walter, Tiemeier H.: Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 24(12):733-6 (2009)
2. R. Battiti: Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks. vol. 5, no. 4, pp. 537-550 (1994)
3. Eysenck, H.J. and Eysenck, S. B.G.: Manual for the Eysenck Personality Inventory. Educational and Industrial Testing Service. (1968)
4. R Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2018)
5. G. Kratzer, R. Furrer, M. Pittavino: Comparison between Suitable Priors for Additive Bayesian Networks. arXiv preprint arXiv:1809.06636 (2018)