Post-doctoral position in statistics
Selection of generalized linear models for hierarchically-structured categorical data: application to the modeling of plant diversity

Duration: 18 months starting during the second semester of 2017.
Eligibility: The candidate shall not have carried out his thesis in Montpellier. The candidate should not have spent more than a year in France over the past 3 years. The candidate must have obtained his PhD in a 6-year period preceding his/her application.
Remuneration according to the CIRAD salary scale (e.g. 28,400 € of net salary per year just after a PhD).

Research Units: Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales (AGAP), CIRAD, Montpellier, France and Institut Montpelliérain Alexander Grothendieck (IMAG), Université de Monpellier, France.
Contact: Yann Guédon (yann.guedon@cirad.fr), Catherine Trottier (catherine.trottier@univ-montp3.fr)

Post-doc program

The objective of the post-doc program will be to develop model selection methods for Partitioned Conditional Generalized Linear Models for categorical data (PCGLM), see Peyhardi *et al*., (2015, 2016). These methods will be applied to a study of rice diversity. In the context of PCGLMs, the hierarchical structure of categories is described by a tree. With a high number of categories, the tree space dimension is naturally huge and this raises new inference questions. In our application context, a partial knowledge about the data structure will be taken into account to constrain this tree space. Model selection will thus mainly entail a tree search in a constrained space and the selection of relevant explanatory variables for each non-terminal vertex of the tree (e.g. the phenotypic traits associated with the separation in sub-species of the cultivated Asian rice).

Post-doc profile

The post-doctoral candidate will have a strong background in statistical modeling, computational statistics and/or applied statistics. In particular knowledge concerning generalized linear models for categorical responses (Agresti, 2013, Tutz, 2011), variable selection and model selection will be very valuable. Some background concerning compiled language such as C++ will be useful beyond knowledge of script language such as R or Python.

References

Agresti A. (2013). *Categorical Data Analysis*, 3rd Edition. Wiley, Hoboken, NJ.
Peyhardi, J., Trottier, C. & Guédon, Y. (2015). A new specification of generalized linear models for categorical data. *Biometrika* 102(4), 889-906.
Peyhardi, J., Trottier, C. & Guédon, Y. (2016). Partitioned conditional generalized linear models for categorical data. *Statistical Modelling* 16(4), 297-321.
Tutz, G. (2011). *Regression for Categorical Data*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.