PhD proposal's title :

Theoretical and practical developments and dissemination of Goodness-of-fit p-values for JSDMs

Main academic fields:

Applied Statistics; Biometrics

Abstract:

Joint-Species Distribution Models (JSDMs) make a lot of mathematical assumptions - sometimes with ecological implications – due to their complex structure and to the numerical limitations required to fit them. Goodness-of-fit p-values are one primary tool used in applied statistics to diagnose parametric models (see state of the art section). New goodness-of-fit p-values, called sampled posterior p-values, have desirable mathematical properties that make them relevant in very different contexts and that give them higher power than more classical GOF p-values which are known to be conservative. Yet, these p-values are currently underused, including in ecology. The aim of the PhD project is to develop further these p-values, both mathematically and "practically", and broadcast their use in ecology with JSDMs as practical cases. More precisely, the objectives of this PhD project are:

(i) To obtain mathematical results on GOF p-values for latent variables and in the case of "external" goodness of fit (where goodness-of-fit is gauged on data not used to estimate the model);

(ii) To implement Monte-Carlo simulation techniques and gather results on how to practically use these p-values, with three directions in mind: use them on controlled sample sizes; better apprehend how results from these p-values can be interpreted (esp. in terms of metric choice and in a context where multiple metrics are used simultaneously); use one p-value vs. a collection of p-values, based on different samples from the estimator.

(iii) Optionally, to disseminate these tools in ecology, with a special emphasis on JSDMs.

At least one scientific manuscript should be targeted on each of these points. It would be welcome to incorporate these tools in statistical decision approaches.

Detailed content:

Context

Although recognized as important tools in applied statistics (cf. state of the art section), goodness-of-fit p-values are not very much used in applied statistics and especially in ecology. Furthermore, it is not well known that there has been a silent clash of paradigms in the development of GOF p-values in the recent past. These tools have furthermore seldom been used in conjunction with model selection tools to help understand in which respects models are better than others and to help better apprehend whether the best model determined by usual approaches match the model selected by the GOF p-values criterion.

At the same time, statistical tools are making more and more numerous and very often uncriticized assumptions to fit models to data. The development of hierarchical models has introduced new layers of parameters, often with very precise assumptions, e.g. about their probability distributions. These assumptions have also been required in fields where fitting the model to data is challenging as is the case for Joint Species distribution models. It is therefore more than welcome to develop, test and calibrate tools of model criticism and GOF p-values are one very interesting candidate to do so. This is the task of the PhD, which will be achieved in a broader French ANR project, called Gambas, to which it will be very much linked.

State of the art

Model criticism is one of the fundamental principles of parametric statistics. Indeed, parametric statistics work through the *a priori* specification of a model, that is then fitted to data. The adequacy of the model with data can be gauged through a post-fit phase of model criticism that allow the user to gauge the confidence he has in the model and the estimates of the model. In applied statistics, Box (1980), McCullagh & Nelder (1989) and Cox (1997) all identified this step as one of the important phases in parametric statistics. In ecology, Hilborn & Mangel (1997) identified model criticism as one of the four main tools of the « ecological detective ». One primary tool for model criticism are GOF p-values. GOF p-values always depend on (i) a metric used to summarize the data and potentially model parameters, called a discrepancy function, and (ii) a method to generate replicate data given the model fit. Based on these two ingredients, it estimates the empirical probability that the discrepancy function on observed data is more extreme than the discrepancy function on replicated data. The analyst has the choice to criticize the fit of the model to data from different points of view, which correspond to different discrepancy functions (see Herpigny & Gosselin, 2015 for an example in ecology). Two kinds of GOF p-values exist: those that use the same data to fit the model and criticize it (which we will call internal GOF p-values) and those that use a separate data set to criticize the model (which we will call external GOF p-values).
A very good synthesis of these tools was made in 2000 by the Journal of the American Statistical Association (Robins et al., 2000). It made it clear that there were numerous different kinds of GOF p-values, and that the most popular ones (called the plug-in p-value and the posterior predictive p-value) were in general conservative. It also made it clear that the (Bayesian) posterior predictive p-value was more conservative than the (frequentist) plug-in p-value. Johnson (2007) latter made it clear that such drawbacks did not hold if the discrepancy function was carefully chosen (but with restrictions on the discrepancy functions that could be used). Other p-values (partial posterior predictive, conditional predictive and the post-processing posterior predictive p-values Hjort et al., 2006) did not have these drawbacks but were numerically very intensive.

An alternative GOF p-value gradually emerged in the last years, which was finally called the sampled posterior p-value (Johnson, 2004, Gosselin, 2011). It was shown mathematically to have asymptotically (in the number of observations) a uniform reference distribution if the model was adequate for whatever discrepancy function (Gosselin, 2011). It was also compared with the plug-in and the posterior predictive p-value on simulated data, to show its greater statistical power to detect discrepancies between the model and data (Gosselin, 2011, Zhang, 2014). This new p-value is a stochastic p-value – associated with the term sampled – since it is based on a random draw of statistical parameters in the posterior distributions of the parameters, which is quite new. This p-value has only been introduced in ecology recently (Herpigny & Gosselin, 2015, Conn et al., 2018).

Yet, transfer to ecology has been rather limited so far. For example, works on Species Distribution models have repeatedly argued against « resubstitution » (Araújo et al., 2005, Araújo & Guisan, 2006, Botkin et al., 2007; see also Rykiel, 1996), that is the use of the same data set to criticize the model and to fit it. This reminds the frequent call by statisticians to avoid using the data twice (e.g. Evans, 1997, Evans, 2000). Furthermore, statistical models in ecology are more and more elaborate, with more and more assumptions being made. Typical assumptions to specify the models concern the probability distribution of the data, the link function used to relate explanatory variables with the mean of the variable to be explained, the assumption or model for the variance or dispersion of the data, assumptions about dependence or independence of the data (e.g. Herpigny & Gosselin, 2015, Saas & Gosselin, 2014). Joint Species Distribution models also introduce assumptions on the way different species are correlated, and often make simplifying assumptions for the probability distribution of data or the link functions (Clark et al., 2017, Wilkinson et al., 2018, Niku et al., 2017). To our knowledge, there has been no development of a procedure to criticize these simplifications in the case of JSDMs.

Description of the content of the PhD

This PhD is decomposed in three main tasks gathered around the development and use of sampled posterior GOF p-values. These tasks give the initial impetus of the PhD; of course, they may evolve during the PhD according to PhD results and ideas and to discussions inside the Gambas project:

Task 1 : Mathematical results : The aim is here to give a more thorough mathematical treatment of sampled posterior p-values than in Gosselin (2011). At least, this would include the adaptation of the procedure to diagnose latent variables in the model and the proposal of similar results for external goodness of fit p-values. Preliminary results are available on these. Potential other mathematical inputs could concern having a more precise mathematical treatment of when the asymptotic behavior is reached, a link of these results with other model evaluation or comparison frameworks or linking these p-values with the field of statistics for decision.

Task 2: Simulation analyses to guide use of sampled posterior p-values: the aim here is to guide the use of these p-values based on simulations or on logical considerations. The principal points on which we intend to work are:

- To use p-values on controlled sample sizes: the problem here is that every model is not a perfect fit to reality. This implies that provided we have a large enough sample size, we should asymptotically reach very low GOF p-values, indicating departures of almost every discrepancy function between the model and the data. The aim of this subtask would be to test the possibility to reduce the sample size – when the sample size is very large – on which the p-value is calculated to have p-values that detect non-minor problems (Gosselin, 2011).
- To better apprehend how results from these p-values can be interpreted, on three directions that are linked: (i) choice to transform (normalize) or not data (Gosselin, 2011); (ii) choice of discrepancy measures to diagnose specific problems in the fit of the model to data. We will here have a specific focus on the main assumptions of JSDMs related to the Gambas project; (iii) way of interpreting “significant” departures between a model and data in a context where multiple metrics are used simultaneously: does the method allow to detect precisely the problem of the model or do we detect many departures simultaneously?
- To devise whether we should use one p-value for which we have mathematical results vs a collection of p-values, based on different samples from the estimator, for which we do not have a precise reference distribution. This subject was already

tackled by Johnson (2004) and Gosselin (2011), but the idea would be to be more specific, especially in terms of the context of application. It is in particular likely that the answer will be different if we use GOF p-values in a model comparison environment or not.

Preliminary results are available on the second and third topic. Some simulations in this task should be very much tied with the simulations being done in the Gambas project.

Task 3: Dissemination of the main results of the PhD in ecology, esp. relative to the use of JSDMs. We have already noticed that sampled GOF p-values are underused and not well known in ecology. The aim of this task is to write a paper that will allow a better acknowledgement of these tools, with JSDMs as an example. The paper should if possible compare this strategy with other tools frequently used by ecologists to evaluate statistical models. This could be done by collecting cases from the ecological community or ecological literature and apply the different tools – including sampled posterior p-values – on these cases. This paper will be accompanied by scripts or libraries (a priori as a package for the R Statistical software) that will make these tools publicly available to potential users. One possibility would be to foster the development of these tools in existing R packages (such as DHARMa).

Tasks 1 & 2 are the core of the PhD. The paper in Task 3 is optional for the PhD student, depending on the time left to do it and on his/her skills and ease with it.

Publications of the team on related subjects:

Gosselin, F., 2011. A New Calibrated Bayesian Internal Goodness-of-Fit Method: Sampled Posterior p-values as Simple and General p-values that Allow Double Use of the Data. Plos One, 6(3), e14770.

Herpigny, B., and Gosselin, F. 2015. Analyzing plant cover class data quantitatively: customized cumulative zero-inflated beta distributions show promising results. Ecological Informatics **26**(3):18-26. doi:10.1016/j.ecoinf.2014.12.002.

Godeau, U., Bouget, C., Piffady, J., Pozzi, T. and Gosselin, F. (Submitted). The importance of being random! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models.

Planning of the PhD

Year 1:
Familiarization/reformulation of the PhD subject; literature reviews; contacts, especially in the Gambas project, esp. on tasks 2 & 3; start of task 1 and task 2; writing a first paper, a priori on Task 1.

Year 2:
Continuing simulations. Writing a second paper, a priori on task 2. Working on R codes or packages for task 3. Attendance to an international conference.

Year 3:
Writing at least a third paper, a priori on task 3. Writing potentially another paper. Writing the PhD.

Organization

PhD Duration:

3 years from approx. October 2020

PhD Director:

Frédéric Gosselin
INRAE
UR EFNO
Domaine des Barres
45290 Nogent sur Vernisson
02 38 95 03 58
frederic.gosselin@inrae.fr

Co-supervision by Camille Coron (Laboratoire de Mathématiques, Université d'Orsay) and Didier Chauveau (Institut Denis Poisson, Université d'Orléans).

Inscription in Orléans University a priori with Didier Chauveau as official director. Ecole Doctorale 551 - Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes (MIPTIS)

The PhD will be located in INRAE, Nogent-sur-Vernisson, a forest ecology lab where the supervisor works. The PhD student will frequently go to Orléans and Orsay University to discuss with Camille Coron and Didier Chauveau and/or attend math seminars. The supervisor has the experience of this kind of PhD, having been located during his PhD in an ecology lab in the South of France and affiliated to a math University in Paris.

Candidates selection

Candidates have to send :

(i) their CV,
(ii) a motivation letter,
(iii) two letters of recommendation,
(iv) a written academic document by the candidate as close as possible to the topic,
(v) official master records.

by e-mail to Frédéric Gosselin (frederic.gosselin@inrae.fr) before **between the 15 June 2020 and the 30 june 2020**.

Hopefully selected candidates will be interviewed end of June and during July by researchers involved in the project.

Funding :

By INRAE through the ANR Gambas project.
Month Salary before social contributions: around 1877€
Month Salary left to the candidate (before revenue tax): a little more than 1500€

Candidate profile

The candidate should have a training in applied mathematics with a strong background in

statistics, probability and statistical computing. He/she should also be interested by applications of mathematical tools to the applied field of ecology.
The candidate should have good skills in scientific English (for reading, speaking and writing) and a good organization of work (both a good autonomy and good reporting skills).

References :

Araújo, M. B., R. G. Pearson, W. Thuiller and M. Erhard, 2005. Validation of species-climate impact models under climate change. Global Change Biology, 11(9), 1504-1513.

Araújo, M. B. and A. Guisan, 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography, 33(10), 1677-1688.

Botkin, D. B., H. Saxe, M. B. Araujo, R. Betts, R. H.W. Bradshaw et al., 2007. Forecasting the effects of global warming on biodiversity. BioScience, 57(3), 227-236.

Box, G. E. P., 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society, Series A, 143(4), 383-430.

Clark, J.S., D. Nemergut, B. Seyednasrollah, P.J. Turner and S. Zhang, 2017. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. Ecological Monographs , 87(1), 34-56.

Conn, P. B., D. S. Johnson, P.J. Williams, S.R. Melin and M.B. Hooten, 2018. A guide to Bayesian model checking for ecologists. Ecological Monographs.

Cox, D. R., 1997. The current position of statistics: A personal view. International Statistical Review, 65(3), 261-290.

Evans, M., 1997. Bayesian inference procedures derived via the concept of relative surprise. Communications in Statistics - Theory and Methods, 26(5), 1125-1143.

Evans, M., 2000. Comments on Asymptotic distribution of P values in composite null models by J. M. Robins, A. van der Vaart and V. Ventura. Journal of the American Statistical Association, 95(452), 1160-1163.

Gosselin, F., 2011. A New Calibrated Bayesian Internal Goodness-of-Fit Method: Sampled Posterior p-values as Simple and General p-values that Allow Double Use of the Data. Plos One, 6(3), e14770.

Herpigny, B. and F. Gosselin, 2015. Analyzing plant cover class data quantitatively: customized cumulative zero-inflated beta distributions show promising results. Ecological Informatics, 26(3), 18-26.

Hilborn, R. and M. Mangel, 1997, The ecological detective: confronting models with data (). Princeton University Press, Princeton (NJ).

Hjort, N. L., F. A. Dahl and G. Hognadottir, 2006. Post-processing posterior predictive p values. Journal of the American Statistical Association, 101(475), 1157-1174.

Johnson, V. E., 2004. A Bayesian chi(2) test for goodness-of-fit. Annals of Statistics, 32(6), 2361-2384.

Johnson, V. E., 2007. Bayesian Model Assessment Using Pivotal Quantities. Bayesian Analysis, 2(4), 719-734.

McCullagh, P. and J. A. Nelder, 1989, Generalized linear methods (). Chapman, London.

Niku, J., D.I. Warton, F.K.C. Hui and S. Taskinen, 2017. Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. Journal of Agricultural, Biological, and Environmental Statistics , 22(4), 498-522.

Robins, J. M., A. van der Vaart and V. Ventura, 2000. Asymptotic distribution of P values in composite null models. Journal of the American Statistical Association, 95(452), 1143-1156.

Rykiel, E. J. Jr., 1996. Testing ecological models: the meaning of validation. Ecological Modelling, 90(3), 229-244.

Saas, Y. and F. Gosselin, 2014. Simulation-based comparative analysis of spatial count regression methods on regularly and irregularly-spaced locations. Ecography, 37(5), 476-489.

Wilkinson, D.P., N. Golding, G. Guillera-Arroita, R. Tingley and M.A. McCarthy, 2018. A comparison of joint species distribution models for presence–absence data. Methods in Ecology and Evolution .

Zhang, J. L., 2014. Comparative investigation of three Bayesian p values. Computational Statistics and Data Analysis, 79, 277-291.