Comparative study of probability models for agriculture field index data

SAGI
Statistics for the Australian Grains Industry
National | North | West | South

Olena Kravchuk[1], John van der Hoek[2]
[1]Biometry Hub, University of Adelaide, [2]School of Mathematics, University of South Australia
olena.kravchuk@adelaide.edu.au

Abstract

This study is motivated by the challenges of the analysis of field indices in agriculture research. The index data we are interested in represents the quotient of two positively correlated random variables, X/(X + Y), but is only observed as a single variate (index), Z. Based on the analysis of that index, important decisions are made about the ranking of agronomic factors and conditions. Here, we are concerned with probability models for such indices, considering and contrasting several choices for X and Y. The choices include: normal, Beta and log-normal distributions, as well as mixtures of either Beta or log-normal. We state here that the mixture of log-normal provides an identifiable and flexible model for at least three components in each X and Y.

Motivation for HI and other indices of the type X/(X+Y)

- Ratio indices are common in agricultural field research practices
- Examples: harvest index (HI), various nutrient use efficiencies (NUE), multiple vegetation indices (including NDVI)
- Harvest index = Commercial Yield/(Commercial Yield + By-Product Yield), or in application to cereals, HI = Grain Yield/Above-ground biomass
- High-throughput phenotyping allows, in principle, direct derivation of HI without measuring grains or biomass
- The current role of indices in plant breeding and agronomics practices is important. Hay (1995) envisaged "*as cropping is forced into marginal, low-yielding areas by population pressures, there will be an increasing need to understand the physiology of cultivars adapted to less favourable environments*"
- Biophysical sink/source models have been proposed recently requiring positive association between biomass and grain yield

Conditions on the support of X and Y and their correlation

- X and Y are non-independent random variables
- Neither X nor Y are observable
- Z = X/(X + Y) = $(1+Y/X)^{-1}$ is observed, i.e. Y/X is observable
- We require r.v. X, Y > 0 (e.g. modelling biomass), and additionally consider corr(X,Y)>0 (e.g. stronger architecture of plant provides higher biomass and supports higher yield)

Probability models

1. X, Y are normal. Drawbacks: can be negative, ratio distribution has a polynomial-tail component
2. X, Y are Beta. Drawbacks: finite support, correlation challenge, problems with parameter estimation
3. X, Y are log-normal. Benefits: strictly positive, exponential tails

X and Y	Theory	Simulations	Interpretation
Normal	Developed	Straightforward	Challenging due to Cauchy in the mixture of the ratio distribution
Beta	Developed	Straightforward	Challenging due to the estimation problems of the joint distribution
Log-normal	Developed	Straightforward	Unambiguous

Mixtures of log-normals

Consider *X* and *Y* both mixtures of components from the same set of log-normal r.vs. *LN* = (LN_1, LN_2, LN_n), $LN_i \sim f_{\mu_i,\sigma_i^2}$.

Can think about *n* hidden factors.

- Let *I* and *J* be component indicators, $I \in \{e_1, e_2, \dots, e_n\}$, $J \in \{e_1, e_2, \dots, e_n\}$, and assign $P(I{=}e_i){=}p_i$, $P(J{=}e_j){=}q_{j.}$ Choosing p_i, $q_{j,}$ the correlation can be of any value in (0,1) [unobserved]
- $X \sim \sum_{i=1}^{n} p_i f_{\mu_i,\sigma_i^2}$,
  $Y \sim \sum_{i=1}^{n} q_i f_{\mu_i,\sigma_i^2}, Y \neq X, \ \frac{Y}{X} \sim \text{mixture of } LN$
- $\frac{Y}{X} = \sum_{ij,i\neq j} I(I = e_i)J(J = e_j)\frac{u_j}{u_i}; u_i \sim LN(\mu_i, \sigma_i^2)$
- Estimating mixtures of log-normal using the EM algorithm. Y/X can be easily simulated.

Ratio of two-component log-normal mixtures with a common component

It can be shown that to meet the constraints for X and Y, the set has to contain at least three variables, *n* ≥ 3.

Consider *n* = 3, with *X* and *Y* being mixtures of one common and one idiosyncratic components each:

X is a mixture of LN_1, LN_2, *Y* is a mixture of LN_2, LN_3

For *n* = 3, the following is estimable

$\mu_2 - \mu_1$	$\sigma_2^2 + \sigma_1^2$
$\mu_3 - \mu_1$	$\sigma_3^2 + \sigma_1^2$
$\mu_3 - \mu_2$	$\sigma_3^2 + \sigma_2^2$

Conclusion and further work

Mixture of log-normal components is a suitable model for field indices of type X/(X+Y), which assume that X and Y positive and positively correlated and are driven by some hidden factors. Mixture modelling is suitable for predictive purposes. Applications to harvest index data from field trials for wheat breeding are under way.

THE UNIVERSITY of ADELAIDE

GRDC
GRAINS RESEARCH & DEVELOPMENT CORPORATION

Government of South Australia
Primary Industries and Regions SA

SARDI
SOUTH AUSTRALIAN RESEARCH AND DEVELOPMENT INSTITUTE