On the quest for distributions, still worth the effort?

Adelchi Azzalini

Università di Padova, Italia

Meeting on 'Statistics & Econometrics'
University of Athens, 15–16 April 2017

Side conditions

- Some personal views,
  with an asymmetric coverage of the available material
- Hence not a review
- Focus is on continuous distributions,
  but some points hold more generally

Distributions, a classical theme

- Development of distributions: an ever-green theme
- In real data, normal-like distributions should be the normality, but in fact non-normal distributions are very . . . 'normal'
- Hence a key target: modelling departures from normality
- Some early contributors (+many others):
  K. Pearson, Fechner, Perozzo, Edgeworth

Some early work on bivariate frequency tables – 1

- Luigi Perozzo 1881–82, studies on marriage age
- frequency table from Italian marriages in 1878–79
- x=(female age), y=(male age)
- bivariate Gaussian fit not satisfactory
- smoothed by "the known method of Wittstein" (see later)
- alternatively, fit a parametric distribution (sketched):

$$\begin{aligned} \text{const} \quad & \times \exp\left(-a_2x^2 \pm a_3x^3 - a_4x^4\right) \\ & \times \exp\left(-a_2'y^2 \pm a_3'y^3 - a_4'y^4\right) \end{aligned}$$

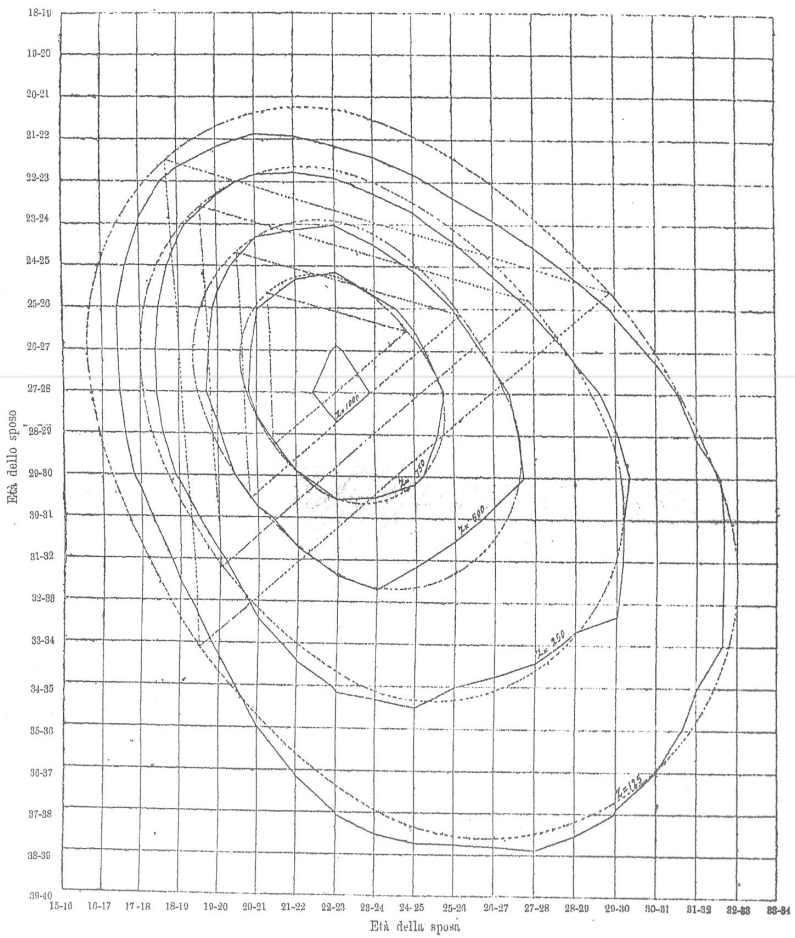
Some early work on bivariate frequency tables – 2

Tav. V.

Matrimoni classificati secondo le varie combinazioni di età degli sposi e delle spose.

Curve di eguale numero di matrimoni nelle varie combinazioni di età dei conjugi.
Le curve sono tracciate in base alle osservazioni dei matrimoni in Italia, avvenuti nel biennio 1878-79, perequate col metodo di Wittstein. I numeri assoluti dei matrimoni effettuatisi sono stati ridotti ad un totale di 100,000 (V. Tavola numerica II).

Scala — cm. 0,9 per ogni anno di età.



A recollection of classical constructions

- Fechner (two-piece distribution with normal components)
- K. Pearson families (via solutions of differential equation)
- normal mixtures – Pearson, again
- Edgeworth generalized law (via suitable expansions)
- transformations of normal variates
- . . . et cetera. . .
- bivariate extensions developed already in early 20th century (see review of Pretorious, 1930)

A little-known early formulation

- F. de Helguero (1908) criticism of most earlier formulations as mathematical constructions which give no clue on the source of non-normality
- Idea of non-normality via a *selection mechanism* :

$$\text{const} \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \{1 - Q(x)\}$$

  where $Q(x)$ is probability of *censoring* a value x
- Tractable case occurs with linear $Q(x)$, i.e. uniform distr'n
- Extend so that $Q(x)$ allows 'thickening' of tails
- Premature death of author left the idea undeveloped

Relatively more recent tools

- Focus on d-dimensional density functions
- From multivariate normal to elliptical families:

$$\text{density at } x = \text{const} \times g\left((x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

- a powerful constructive mechanism: *copulæ* wildly used, allow to combine freely marginals and dependence (note 'wildly')
- another general formulation: 'symmetry modulation'

Symmetry modulation – basics

- symmetry-modulated distr's = skew-symmetric distr's
- for any d-dimensional density $f(x)$ can write

$$f(x) = 2\, f_0(x)\, G(x)$$

where $f_0(x) = f_0(-x)$ and

$$G(x) \geq 0, \quad G(x) + G(-x) = 1 \qquad (x \in \mathbb{R}^d)$$

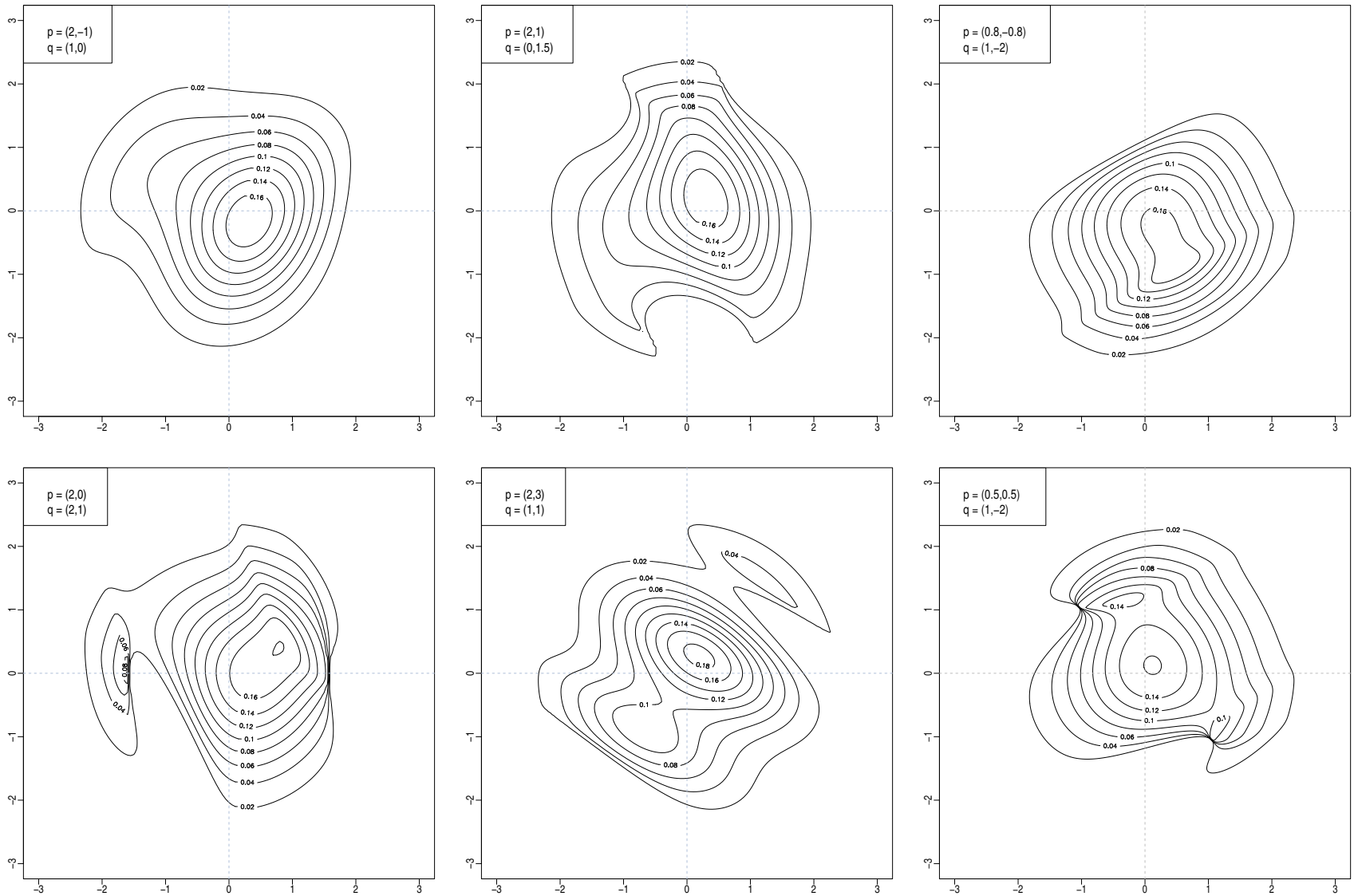
- usually, employed by selecting f_0 and $G(x)$ to construct $f(x)$: start from 'base' $f_0(x)$ and modulate it with

$$G(x) = G_0\{w(x)\}$$

where G_0 is symmetric CDF and $w(-x) = -w(x)$

- an underlying selection mechanism on f_0 regulated by $G(x)$
- this provides a stochastic representation, from here obtain useful properties
- comprehensive account by Azzalini & Capitanio (2014)

Examples: modulation of bivariate standard normal



Symmetry modulation – connections

When 'base' density f_0 is normal,

- includes de Helguero's (1908) idea;
- similarly, connection with Heckman selection model (1976);
- it includes classical formulation for 'stochastic frontier' (SFA):

$$y = \mu(\text{production factors}) + \underbrace{\sigma_1\,\varepsilon_1 - \sigma_2|\,\varepsilon_2|}_{\textit{asymmetric}}$$

These connections allow to extend existing formulations, e.g.

- Marchenko & Genton (2012), 'robustified' Heckman model
- Azzalini, Kim & Kim (2016), Heckman model for GLMs
- Tancredi (2002), SFA with Student's random components

Plenty of resources

- Vast repertoires of distributions are available
- Ease-to-use techniques exist to design new distributions
- Do we need to invest more effort here?

Any problem?

- "Recipe for Disaster: The formula that killed Wall Street"

  Here's what killed your 401(k). David X. Li's Gaussian copula function as first published in 2000. Investors exploited it as a quick—and fatally flawed—way to assess risk.

  (Wired Magazine, March 2009)

- Paul Embrechts (2009): "For me, this is akin to blaming Einstein's $E = mc^2$ formula for the destruction wreaked by the atomic bomb"
- David X. Li (2005): "Very few people understand the essence of the model."
- misuse (of the copula idea, in this case) is to be blamed
  . . . but how to help preventing misuse?

Statistical modelling

Choice of distribution is a key part of the modelling process

Cox (1997), 'Desiderata for a probabilistic model':

1. the model should establish a link with underlying substantive knowledge or theory;
2. the model should allow comparisons with previous related studies of the topic;
3. the model should be consistent with or suggest a possible process that might have generated the data;

(...)

6. the fit to data should be adequate.

Choice of distribution in statistical modelling

- subject-matter knowledge should be taken into account
- this is especially important when models/distributions are to be used outside the domain of the data used for fitting
- interpretability of the formulation is key for successful cooperation with applied environment
- theoretical work should promote above modelling criteria, not only data fitting and flexibility of distributions
- an implication is effort to better understand formal properties with respect to potential applications

References

- Azzalini & Capitanio (2014), monograph, Cambridge UP
- Azzalini, HM Kim & HJ Kim (2016), under reviison
- Cox (1997), Int. Stat. Review, 65, 261-290
- de Helguero (1908), IV Internat. Congress Mathematicians
- Marchenko & Genton (2012), JASA 107, 304–317.
- Perozzo (1881–82), Atti R. Accademia dei Lincei
- Pretorious (1930), Biometrika, 22, 109–223
- Tancredi (2002), Tech.report, Univ.Padua, Italy