One-day course
on symmetry-modulated distributions

Adelchi Azzalini

Università {degli Studi} di Padova

Motivation

Do we need more distributions?

- An arsenal of probability distributions is available
- Classical constructions
  by K.Pearson, Fechner, Edgeworth, Johnson, Burr, etc.
- Two popular 'general' approaches:
  - copulae
  - symmetry-modulated distributions,
    AKA skew-symmetric distributions

- Question: do we need more distributions?

Illustration: QQ-normal probability plots from two samples

Normal QQplot, sample size = 30

Normal QQplot, sample size = 3000
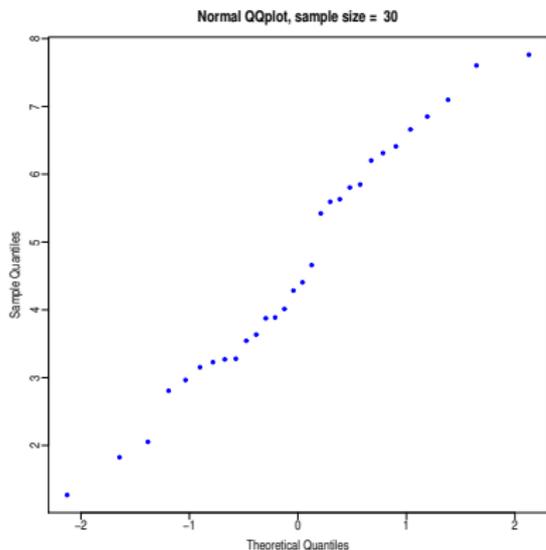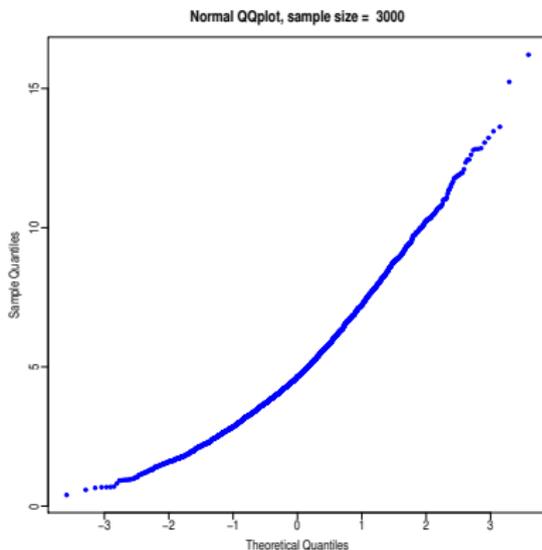
$n = 30$ $n = 3000$

```
# R code for the first plot, except title line
set.seed(1)
n0 <- 30    # for the second plot set n0 <- 3000
x0 <- rgamma(n0, 5, 1)
qqnorm(x0, pch=20, col="blue")
```

Larger datasets require more accurate modelling

- The two datasets are sampled from the same distribution
- yet the two plots deliver disparate messages!
- Moral: the sample size matters.

- Nowadays datasets get larger, larger, larger
- More data is welcome, but also challenging
- We need flexible tools for accurate modelling of large datasets
- skewness is the 'first step' moving away from normality

Multivariate datasets are increasingly more frequent

- data collection is more often multivariate, possibly highly so
- many classical formulations are univariate
- special interest in developing
  - multivariate distributions
  - flexible multivariate distributions
  - flexible, yet mathematically tractable, multivariate distributions

Our plan of work

A tutorial to symmetry-modulated distributions:

- introduce main concepts in the univariate case
- focus on key special cases
- extend concepts to the multivariate settings
- sketch of some extensions

Motivation
ooooooo

Basics (d=1)
●ooooooooo

Plus (d = 1)
ooooooo

Data
oooooooooo

Basics (d > 1)
ooooooooooooo

Extra
ooo

Closing
ooo

Basics (d=1)

Skew-normal distribution – idea

Idea: start from a normal distribution and 'perturb' it.
Perturbation, or modulation, is achieved by a selection mechanism.
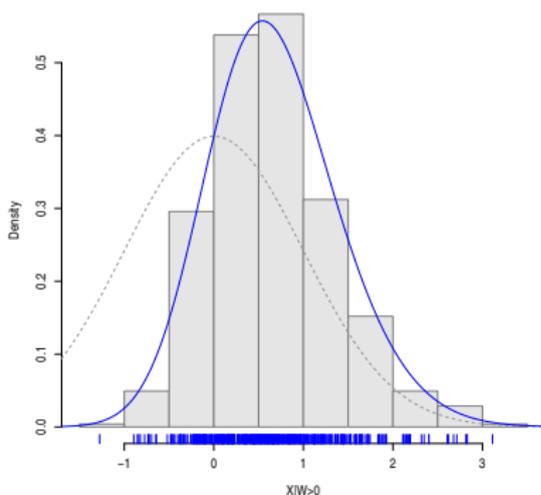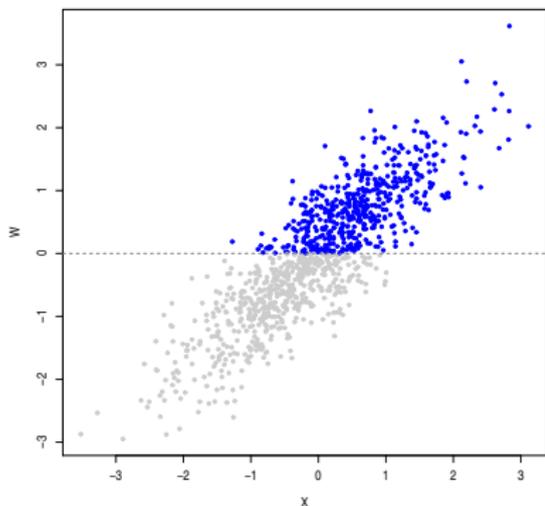
This is the basic idea for one route, but other routes exist.
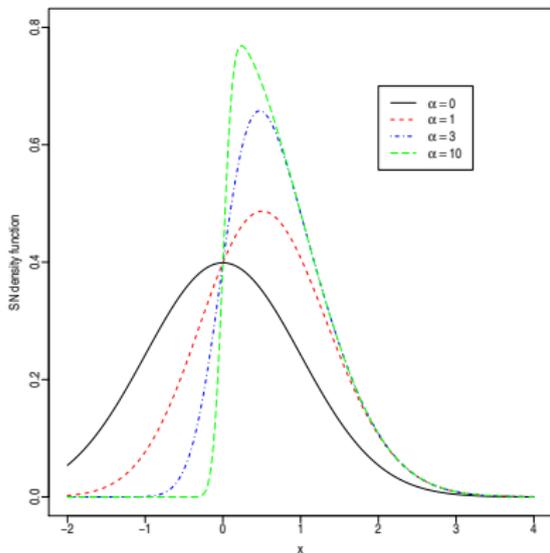
Skew-normal distribution – compute density function

$$\text{assume} \quad : \quad (X, W) \sim N_2(0, \Sigma), \qquad \Sigma = \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix}$$

$$\text{recall} \quad : \quad (W|X = x) \sim N(\delta x, 1 - \delta^2)$$

$$(\text{density at } x|W \geq 0) \quad = \quad \frac{1}{dx} \mathbb{P}\{X \in (x, x + dx)|W \geq 0\}$$

$$= \quad \frac{1}{dx} \frac{\mathbb{P}\{X \in (x, x + dx) \cap W \geq 0\}}{\mathbb{P}\{W \geq 0\}}$$

$$= \quad \frac{1}{dx} \frac{\mathbb{P}\{X \in (x, x + dx)\} \mathbb{P}\{W \geq 0|X = x\}}{1/2}$$

$$= \quad \boxed{2\, \varphi(x)\, \Phi(\alpha x),} \qquad \alpha = \frac{\delta}{\sqrt{1 - \delta^2}} \in \mathbb{R}$$

$$\text{write} \quad : \quad Z \equiv (X|W \geq 0) \sim \mathrm{SN}(\alpha)$$

Skew-normal distribution – density function plots

$\alpha > 0$: positive asymmetry
$\alpha = 0$: null asymmetry, i.e. $N(0, 1)$
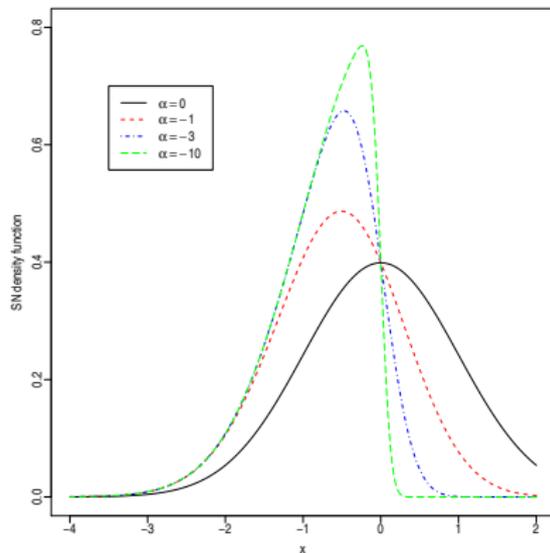
$\alpha < 0$: negative asymmetry
$\alpha = 0$: null asymmetry, i.e. $N(0, 1)$

Cautionary note

Beware of confusing terminology:

after 1985 the term 'skew-normal' has been adopted for other distributions

A general result

> ### Lemma (Univariate version)
>
> If f_0 is PDF and G_0 a continuous CDF on $\mathbb{R}$, both symmetric about 0, then
> $$f(x) = 2\, f_0(x)\, G_0\{w(x)\}, \qquad x \in \mathbb{R},$$
> is a proper density function for any odd function w.

Proof. Denote $X \sim f_0$ and $T \sim G_0$, independent rv's.
The distribution of $T - w(X)$ is symmetric about 0.
Then

$$
\begin{aligned}
\tfrac{1}{2} &= \mathbb{P}\{T - w(X) \leq 0\} \\
&= \mathbb{E}_X\{\mathbb{P}\{T \leq w(x)|X = x\}\} \\
&= \int_{\mathbb{R}} G_0\{w(x)\}\, f_0(x)\, \mathrm{d}x
\end{aligned}
$$

Some comments

- Above result allows to combine freely f_0, G_0 and w:
  a huge variety of constructions are possible

- need to select the useful ones

- Various extensions of the lemma:
  multivariate, discrete variables, bounded support, non-odd w
  . . . .

- We can equivalently work with

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

  which satisfies

$$G(x) \geq 0, \qquad G(x) + G(-x) = 1\,.$$

Random rumber generation / stochastic representation

Crude version Generate $X \sim f_0$ and $T \sim G_0$ independently and set

$$Z = (X | T \leq w(X))$$

Drawback: reject sampled values with $T > w(X)$,
half of the total number on average.

Improved version

$$\boxed{Z = \begin{cases} X & \text{if } T \leq w(X) \\ -X & \text{otherwise} \end{cases}}$$

No rejection of sampled values

Perturbation invariance

- Recall stochastic representation

$$Z = \begin{cases} X & \text{if } T \leq w(X) \\ -X & \text{otherwise} \end{cases}$$

- then $|Z|$ is distributed like $|X|$, write $|Z| \overset{d}{=} |X|$

- more generally: $\boxed{t(Z) \overset{d}{=} t(X)}$ for any even $t(\cdot)$
  $\Rightarrow$ property of perturbation (or modulation) invariance

- Example: if $Z \sim \mathrm{SN}(\alpha)$, then $Z^2 \sim \chi_1^2$

Motivation
0000000

Basics (d=1)
000000000

Plus ($d = 1$)
●000000

Data
0000000000

Basics ($d > 1$)
00000000000

Extra
000

Closing
000

Plus ($d = 1$)

More on SN: other stochastic representations

Representation by conditioning/selection
this was how we introduced the SN distribution

Additive representation

- If U_0, U_1 are independent $N(0,1)$ variables, then
$$\boxed{Z = \sqrt{1 - \delta^2}\, U_0 + \delta\, |U_1| \sim \mathrm{SN}(\alpha)}$$
- much used to develop EM-type algoritms

Representation via minima/maxima

- assume (X, Y) is bivariate standard Normal with $\mathrm{corr}\{X, Y\} = \rho$
- write $\alpha = \sqrt{(1 - \rho)/(1 + \rho)}$
- then $\boxed{\max(X, Y) \sim \mathrm{SN}(\alpha)}$ and $\boxed{\min(X, Y) \sim \mathrm{SN}(-\alpha)}$

More on SN: some formal properties

- Moment generating function has a simple expression:

$$M(t) = 2 \, \exp(\tfrac{1}{2} t^2) \, \Phi(\delta t)$$

$\Longrightarrow$ compute moments

(only odd moments really need computing)

$\Longrightarrow$ derive futher properties

e.g. if $Z \sim \mathrm{SN}(\alpha) \perp\!\!\!\perp U \sim \mathrm{N}(0,1), \quad Z + U \sim \sqrt{2} \times \mathrm{SN}(\tilde{\alpha})$

- Distribution function has a tractable expression

Thick tails

- In many applications we need thicker-than-normal tails
- This feature cannot be achieved by perturbation of $\mathrm{N}(0,1)$
- We must start from a baseline density f_0 in
$$f(x) = 2\, f_0(x)\, G_0\{w(x)\}$$
which already has thick tails
- Preference for those where f_0 allows a tail-regulation parameter

Skew-t (ST) distribution – genesis

- A good choice for f_0 is the Student's t density: $t(x; \nu)$, $\nu > 0$
- Even then, still many possible options, such as the 'linear form'

$$2\, t(x; \nu)\, T(\alpha\, x; \nu)$$

- There are strong reasons for picking up another option
- Recall origin of classical Student's t:

$$Z \sim \mathrm{N}(0, 1) \perp\!\!\!\perp W_\nu \sim \chi^2_\nu \quad \Longrightarrow \quad \boxed{\dfrac{Z}{\sqrt{W_\nu/\nu}} \sim t(x; \nu)}$$

- Use the the same construction with $Z \sim \mathrm{SN}(\alpha)$
  $\Longrightarrow$ obtain the $\boxed{\mathrm{ST}(\alpha, \nu) \textit{ distribution}}$
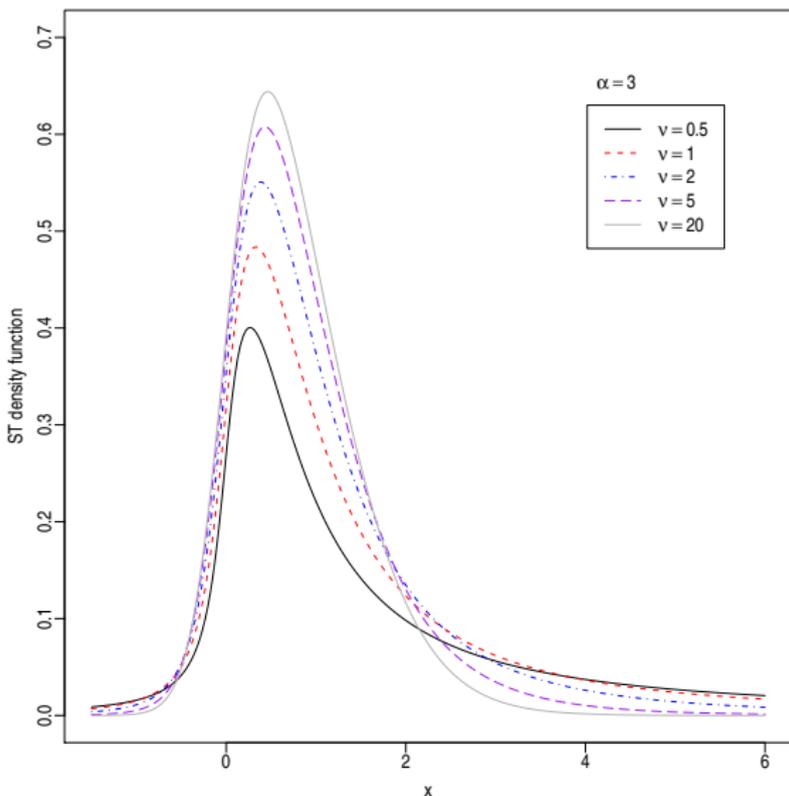- Beware: in literature various other proposals named 'skew-t'

Skew-t (ST) distribution – a closer look

- Algebraic work leads to $\mathrm{ST}(\alpha, \nu)$ density function:

$$2 \, t(x; \nu) \, T\left(\alpha x \sqrt{\frac{\nu + 1}{\nu + x^2}} ; \nu + 1 \right)$$

- $Z \sim \mathrm{ST}(\alpha, \nu) \implies Z^2 \sim F(1, \nu)$
- m-th order moment exist if $m < \nu$, like regular t
- a very wide range of γ_1 (skewness) and γ_2 (excess kurtosis)
  $-\infty < \gamma_1 < \infty, \qquad 0 \leq \gamma_2 < \infty \quad$ (but no $\gamma_2 < 0$)
- widely flexible shape, well-suited for data fitting
  (when complemented with location and scale parameters)
- as $\nu \to \infty$, density $\mathrm{ST}(\alpha, \nu) \to \mathrm{SN}(\alpha)$

Skew-t (ST) distribution – examples of density

Motivation
oooooo

Basics (d=1)
ooooooooo

Plus ($d = 1$)
ooooooo

Data
●ooooooooo

Basics ($d > 1$)
ooooooooooo

Extra
ooo

Closing
ooo

Data

Motivation
oooooo

Basics (d=1)
ooooooooo

Plus (d = 1)
ooooooo

Data
o●oooooooo

Basics (d > 1)
oooooooooooo

Extra
ooo

Closing
ooo

Location and scale parameter

- Let Z be a SN or ST or something-of-the-kind random variable
- For applied work, introduce location and scale parameters:

$$\boxed{Y = \xi + \omega\, Z,} \qquad \xi \in \mathbb{R}, \quad \omega \in \mathbb{R}^+$$

- correspondingly extend our notation to
  $Y \sim \mathrm{SN}(\xi, \omega, \alpha)$ and $Y \sim \mathrm{ST}(\xi, \omega, \alpha, \nu)$
- Note: ξ is not the mean, ω is not the standard deviation

Motivation
oooooo

Basics (d=1)
ooooooooo

Plus (d = 1)
ooooooo

Data
oooooooooo

Basics (d > 1)
oooooooooooo

Extra
ooo

Closing
ooo

Fitting a SN distribution

- Start from simple case of i.i.d. observations $y = (y_1, \ldots, y_n)$
- log-likelihood for SN:

$$\log L(\xi, \omega, \alpha) = \text{constant} - \tfrac{1}{2} n \log \omega - \tfrac{1}{2} \sum_i z_i^2 +$$

$$+ \sum_i \log \Phi(\alpha \, z_i)$$

$$\text{having set} \qquad z_i = (y_i - \xi)/\omega$$

- In a regression model, location depends on covariates x_i, typically in a linear form:

$$\boxed{\xi_i = x_i^\top \beta} \qquad x_i, \beta \in \mathbb{R}^p, \quad i = 1, \ldots, n$$

- log-likelihood $\log L(\beta, \omega, \alpha)$ is as before, except that now

$$z_i = (y_i - \xi_i)/\omega = (y_i - x_i^\top \beta)/\omega$$

Motivation
oooooo

Basics (d=1)
ooooooooo

Plus (d = 1)
ooooooo

Data
oooo●oooooo

Basics (d > 1)
ooooooooooo

Extra
ooo

Closing
ooo

Illustration: fitting SN to phenols content in Barolo wine

$n = 59$

$\hat{\gamma}_1 = 0.8$
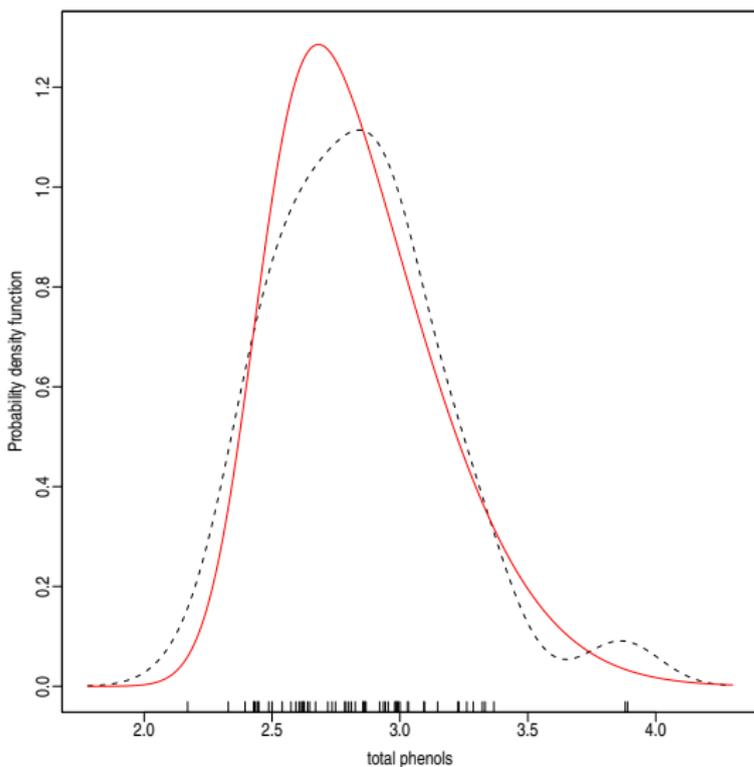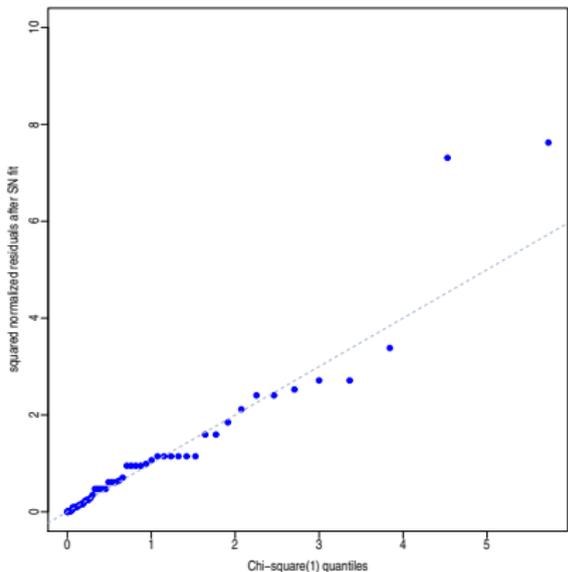
$\dfrac{\hat{\gamma}_1}{\text{std.err.}} = 2.5$

Motivation
oooooo

Basics (d=1)
ooooooooo

Plus ($d = 1$)
ooooooo

Data
ooooo●oooooo

Basics ($d > 1$)
ooooooooooooo

Extra
ooo

Closing
ooo

Illustration: graphical diagnostics of SN fitting

recall : $Z^2 = (Y - \xi)^2/\omega^2 \sim \chi_1^2$

approx : $\hat{Z}^2 = (Y - \hat{\xi})^2/\hat{\omega}^2 \overset{\cdot}{\sim} \chi_1^2$

QQ-plot : $\hat{z}_{(i)}^2$ vs χ_1^2 quantiles

with ST : replace χ_1^2 with $F(1, \hat{\nu})$

SN log-likelihood: some unusual aspects

Two sort of noteworthy phenomena

- 'Transient' sort of occasional events
  Usually with small n, sporadic if n beyond a few dozens
  Similar behaviour fairly common also with other models
  - multiple local maxima
  - $\max \log L$ occurs at $\alpha \to \pm\infty$

- 'Persistent (but local)' behaviour:
  that is, for all samples, but only at $\alpha = 0$
  - stationarity of $\log L$ at point $\alpha = 0$
  - correspondingly, singularity of the information matrix

SN log-likelihood: stationarity of $\log L$ at $\alpha = 0$

$$\text{deviance (LRT)} \quad : \quad D(\theta) = 2\{\log L(\hat{\theta}) - \log L(\theta)\}$$

$$\text{profile deviance} \quad : \quad D(\theta) = 2\{\log L(\hat{\theta}, \hat{\psi}) - \log L(\theta, \hat{\psi}(\theta))\}$$
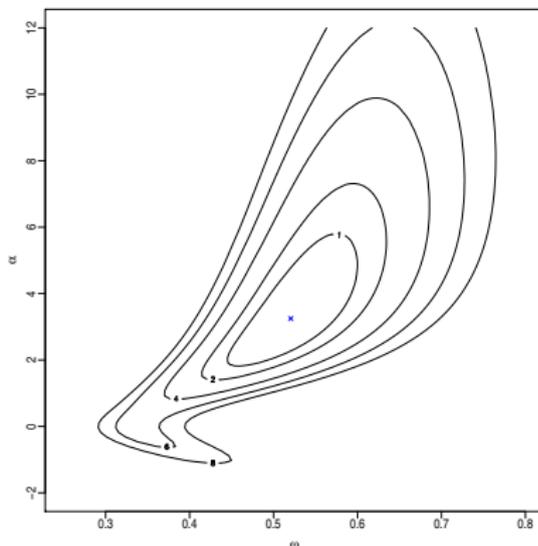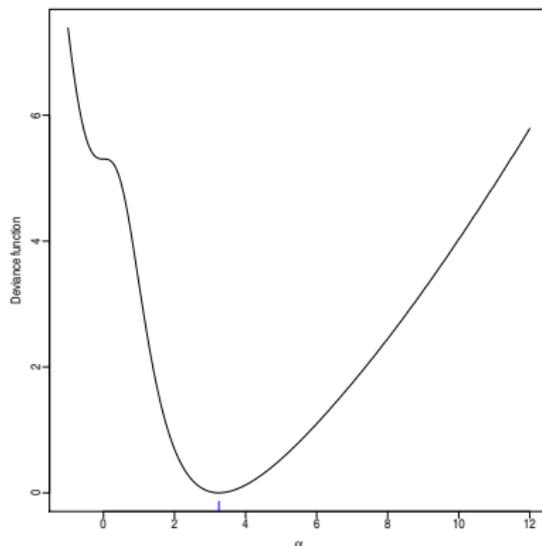
CP for SN

- The twists of $\log L$ at $\alpha = 0$ can be fixed by switching from 'direct' (DP) to 'centred parameterization' (CP)
- Conceptually, we re-parameterize as

$$Y = \xi + \omega\, Z = \mu + \sigma\, Z_0$$
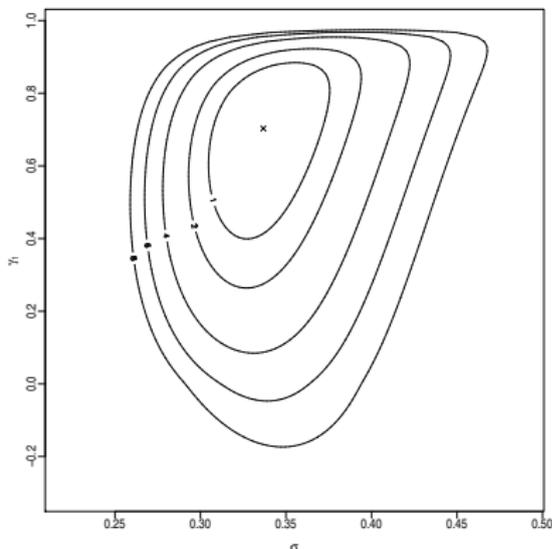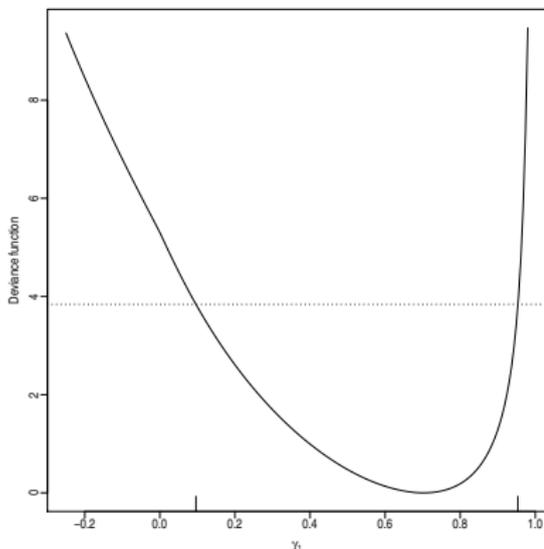
via the 'centred variable'

$$Z_0 = (Z - \mathbb{E}\{Z\})/\mathrm{std.dev.}(Z)$$

- $\mathrm{CP} = (\mu, \sigma, \gamma_1)$
- In parallel, CP avoids singularity of the information matrix
- Importantly, CP is easier to interpret than DP

Motivation
oooooo

Basics (d=1)
ooooooooo

Plus (d = 1)
ooooooo

Data
ooooooooo●o

Basics (d > 1)
ooooooooooo

Extra
ooo

Closing
ooo

SN log-likelihood: using CP with the Barolo data

deviance (LRT) : $D(\theta) = 2\{\log L(\hat{\theta}) - \log L(\theta)\}$

profile deviance : $D(\theta) = 2\{\log L(\hat{\theta}, \hat{\psi}) - \log L(\theta, \hat{\psi}(\theta))\}$

ST log L

- With ST model no stationarity of log L at $\alpha = 0$
- hence no singularity of information matrix at $\alpha = 0$
- in fact, these issues are specific 'only' of φ baseline
- still CP useful for easier interpretability

Motivation
ooooooo

Basics (d=1)
oooooooooo

Plus ($d = 1$)
ooooooo

Data
ooooooooooo

Basics ($d > 1$)
●oooooooooo

Extra
ooo

Closing
ooo

Basics ($d \geq 1$)

Multivariate skew-normal distribution: genesis

- SN was constructed from bivariate standard normal (X, W) as

$$Z = (X | W \geq 0)$$

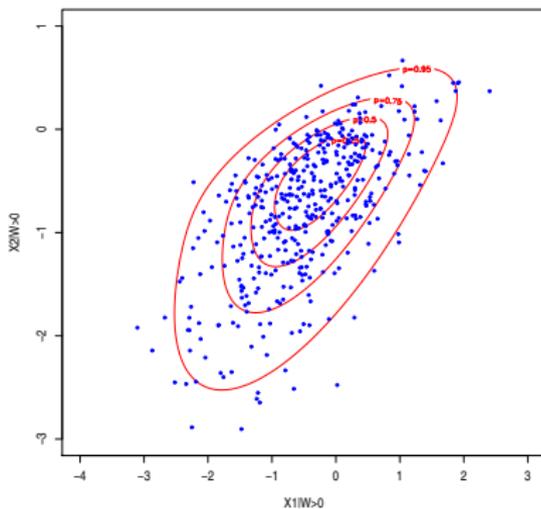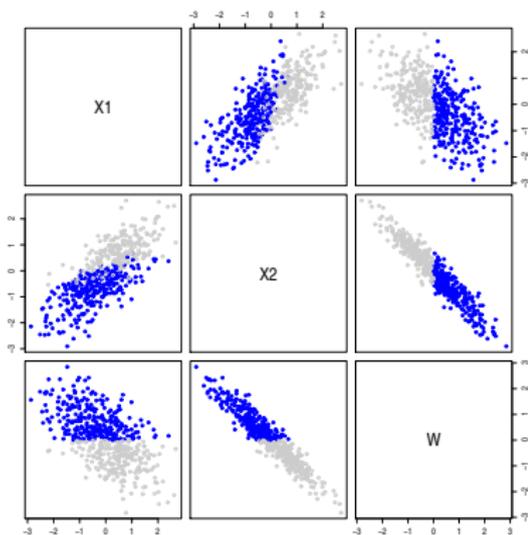- Now start from $(d+1)$-dimensional Normal with std marginals

$$\begin{matrix} d \\ 1 \end{matrix} \begin{pmatrix} X \\ W \end{pmatrix} \sim \mathrm{N}_{d+1}(0, \bar{\bar{\Sigma}})$$

where $\bar{\bar{\Sigma}}$ is a correlation matrix

$$\bar{\bar{\Sigma}} = \begin{pmatrix} \bar{\Omega} & \delta \\ \delta^\top & 1 \end{pmatrix}$$

- and then use the same conditioning process: $Z = (X | W \geq 0)$ except that now X is d-dimensional

Multivariate SN – illustration of genesis

Multivariate SN — basic formal facts

- If $Z = (X|W \geq 0)$, its density function turns out to be:

$$\boxed{2\,\varphi_d(x; \bar{\Omega})\,\Phi(\alpha^\top x),} \qquad x \in \mathbb{R}^d,$$

where $\varphi_d(x; V)$ is $\mathrm{N}_d(0, V)$ density and

$$\alpha = \left(1 - \delta^\top \bar{\Omega}^{-1}\delta\right)^{-1/2} \bar{\Omega}^{-1}\delta \quad \in \mathbb{R}^d$$

- Moment generating function has a simple expression
- Additive representation extends to multivariate SN:

$$\boxed{Z = \left(I_d - \mathrm{diag}(\delta)^2\right)^{1/2} U_0 + \delta\,|U_1|}$$

where $U_0 \sim N_d(0, \Psi) \perp\!\!\!\perp U_1 \sim \mathrm{N}(0, 1)$.

Multivariate SN — include location and scale

- Start from $Z = (Z_1, \ldots, Z_d)^\top$ with density $2\, \varphi_d(x; \bar{\Omega})\, \Phi(\alpha^\top x)$
- introduce location and scale:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_d \end{pmatrix} + \begin{pmatrix} \omega_1 & & 0 \\ & \ddots & \\ 0 & & \omega_d \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix}$$

- write more compactly

$$\boxed{Y = \xi + \omega\, Z}$$

where $\omega = \mathrm{diag}(\omega_1, \ldots, \omega_d)$

- notation: $\boxed{Y \sim \mathrm{SN}_d(\xi, \Omega, \alpha)}$ where $\Omega = \omega\, \bar{\Omega}\, \omega$
- density at $x \in \mathbb{R}^d$:

$$\boxed{2\, \varphi_d(x - \xi; \Omega)\, \Phi(\alpha\, \omega^{-1}(x - \xi))}$$

Recall elliptical families

- Recall continuous elliptically contoured (EC) distributions
- Density constant on ellipsoids:

$$f(x) = \frac{c_d}{(\det \Sigma)^{1/2}} \, g_d \left((x - \mu)^\top \Sigma^{-1} (x - \mu) \right), \qquad x \in \mathbb{R}^d$$

- Notation: $X \sim \mathrm{EC}_d(\mu, \Sigma, g_d)$
- density is centrally symmetric about μ: $f(x - \mu) = f(\mu - x)$
- Extends the normal distribution which corresponds to

$$g_d(u) = \exp(-u/2)$$

- The key aspect is that the EC family encompasses many others
- and it still preserves various properties of normal distribution:
  - family closed under marginalization
  - family closed under conditioning
  - conditional mean is linear function of the conditioning variables

Skew-elliptical distributions

- Start from

$$\begin{pmatrix} X \\ W \end{pmatrix} \sim \mathrm{EC}_{d+1}(0, \bar{\Sigma}, g_{d+1})$$

- and apply the 'usual' conditioning (or selection) process:

$$Z = (X|W > 0)$$

- Introduce location and scale: $Y = \xi + \omega\, Z$
- Terminology: Y and Z have skew-elliptical distribution (SEC)
- If (X, W) is normal, reproduce $Y \sim \mathrm{SN}_d(\xi, \Omega, \alpha)$
- Another noteworthy case with $(X, W) \sim t_{d+1}(0, \bar{\Sigma}, \nu)$:

$$Y \sim \mathrm{ST}_d(\xi, \Omega, \alpha, \nu)$$

- density of normalized r.v. $Z \sim \mathrm{ST}_d(0, \bar{\Omega}, \alpha, \nu)$:

$$2\, t_d(z; \bar{\Omega})\, T\left(\alpha^\top z \sqrt{\frac{\nu + d}{\nu + z^\top \bar{\Omega}^{-1} z}}; \nu + d \right), \qquad z \in \mathbb{R}^d$$

A general result

> ### Lemma (Multivariate version)
>
> If f_0 is a PDF on $\mathbb{R}^d$ and G_0 a continuous CDF on $\mathbb{R}$, both symmetric about 0, then
>
> $$f(x) = 2 f_0(x) G_0\{w(x)\}, \qquad x \in \mathbb{R}^d,$$
>
> is a proper density function for any odd function $w(\cdot)$ on $\mathbb{R}^d$.

Proof: a simple extension of the univariate version.

Notes:
(1) f_0 symmetric on $\mathbb{R}^d$ means $f_0(x) = f_0(-x)$ for all $x \in \mathbb{R}^d$
(2) w odd function on $\mathbb{R}^d$ means $w(-x) = -w(x)$ for all $x \in \mathbb{R}^d$.

A general result — comments

- Both SN_d and ST_d have density like $f(x)$ in the lemma
- Can show that all SEC distributions have this structure with 'baseline density' f_0 of elliptical type
- But the lemma allows far more general constructions, unrelated to SEC's
- This modulation process can produce all sort of shapes, even quite bizarre ones, not just 'skew'
- Next plots illustrate this point using

$$f_0 = \varphi_2, \qquad G_0 = \Phi$$

$$w(x_1, x_2) = a_1 x_1 + a_2 x_2 + a_3 x_1^3 + a_4 x_2^3 + a_5 x_1^2 x_2 + a_6 x_1 x_2^2$$

Examples of modulated bivariate normal densities
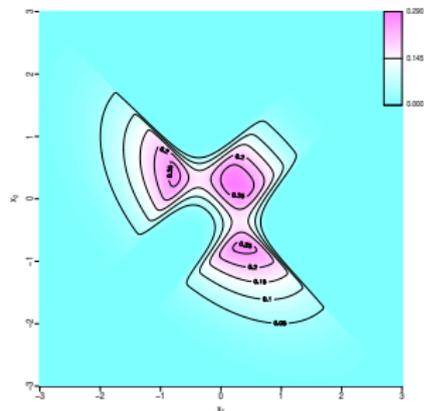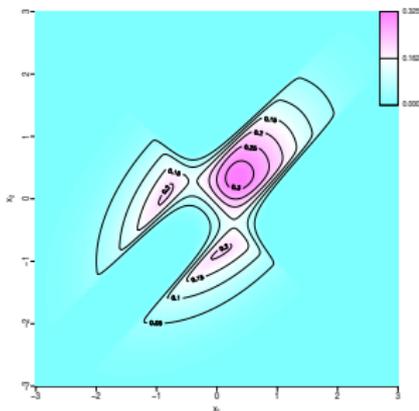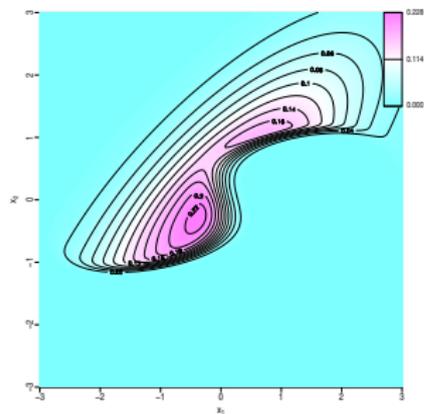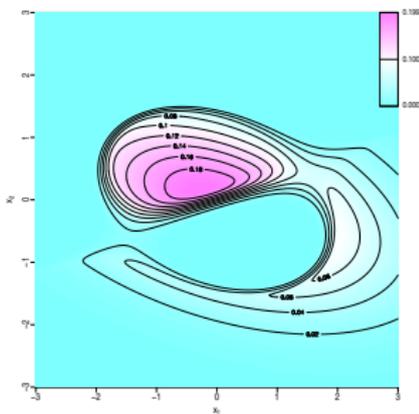
Some formal properties of the general construction

$$f(x) = 2\, f_0(x)\, G_0\{w(x)\}, \qquad x \in \mathbb{R}^d$$

Stochastic representation If $X \sim f_0 \perp\!\!\!\perp T \sim G_0$, then

$$Z = \begin{cases} X & \text{if } T \le w(X) \\ -X & \text{otherwise} \end{cases} \quad \text{has density } f(\cdot)$$

Perturbation (or modulation) invariance Now holds multivariate:

$$t(Z) \stackrel{d}{=} t(X)$$

for any even $t(x)$, mapping $\mathbb{R}^d \to \mathbb{R}^q$

Motivation
○○○○○○

Basics ($d=1$)
○○○○○○○○○

Plus ($d = 1$)
○○○○○○○

Data
○○○○○○○○○○

Basics ($d > 1$)
○○○○○○○○○○○

Extra
●○○

Closing
○○○

Extra

Many additional developments

- Many forms of generalization exist
- The more tractable case is the extended SN and alike:
  start from $(X, W) \sim \mathrm{N}_2$ and take $(X|W \geq c)$ with $c \in \mathbb{R}$
- Important extension: m-dimensional conditioning variable W

  relatively tractable in normal context (Closed SN)

  to some extent also tractable in EC class
- General selection mechanism:
  replace $(\cdots|W \geq 0)$ by $(\cdots|W \in C)$ with $C \subset \mathbb{R}^m$
  (For general C, difficult to find normalizing constant)

Use in statistical methods and applied areas

Two intersecting levels of work:

- Extensions of standard statistical methods
- Application in diverse fields, often with suitable methodological adaption of existing techniques

Many domains:

- classical areas of statistical methods, such as longitudinal data, factor analysis, item response analysis, . . .
- much impact especially in model-based clustering
- flexible distributions provide a route to robustness
- but also in environmental risk, medical statistics, finance, econometrics, income distribution, data confidentiality, insurance, industrial statistics and reliability, cell biology, forestry, *et cetera*. . .

Motivation
oooooo

Basics (d=1)
ooooooooo

Plus ($d = 1$)
ooooooo

Data
ooooooooooo

Basics ($d > 1$)
oooooooooooo

Extra
ooo

Closing
●oo

Closing

Any future?

- Formidable work has been deployed, but still room for progress
- Extension of standard statistical methods for more flexible models, with applications
- Futher advances possible in the study of flexible distributions

Resources

A complete list of references would take many pages.
An absolutely minimal list is:

- MG Genton (2004), edited volume, C&H/CRC
- A Azzalini & A Capitanio (2014), monograph, Cambridge UP

- R software: `https://cran.r-project.org/package=sn`