More on multivariate distributions

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

More on multivariate distributions

Outline

Location-scale families

Recall that the n-variate normal $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be decomposed as

$$\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{H}\boldsymbol{Z}, \quad \text{with } \mathbf{H} \text{ a } n \times n \text{ matrix s.t. } \mathbf{H}\mathbf{H}' = \boldsymbol{\Sigma},$$

where $\boldsymbol{Z}$ is a vector of n independent standard normals.

If $\boldsymbol{Z}$ is not normal, but has zero mean and uncorrelated elements, we still have $\mathrm{E}\left(\boldsymbol{X}\right) = \boldsymbol{\mu}$ and $\mathrm{Cov}\left(\boldsymbol{X}\right) = \mathbf{H}\mathbf{H}' = \boldsymbol{\Sigma}$, leading to

Family Name: Location-scale (multivariate)

Parameterization	$\boldsymbol{\mu} \in \mathbb{R}^n$, Σ pos. def., g multivariate pdf
Density Definition	$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{\lvert\boldsymbol{\Sigma}\rvert}} g\left(\boldsymbol{\Sigma}^{-0.5}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)\right)$
Moments	$\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ (g is standardized with finite variance)

It is often convenient to pick g such that it is the density of a vector of independent standardized random variables.

Elliptical distributions

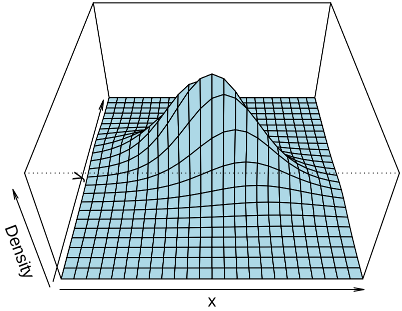
One interesting subclass of multivariate location-scale distributions of the class of the elliptical distributions, defined as

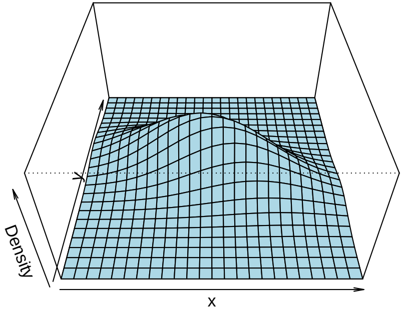
Family Name: Elliptical distributions

Parameterization	$\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ pos. def., $g : g\left(x^2\right)$ integrable
Density kernel	$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto g\left((\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$

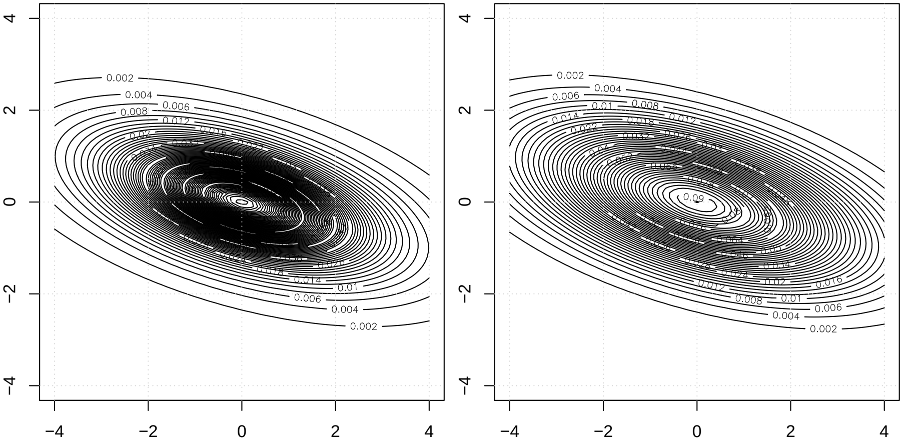
- The name comes from the fact that the level curves of the density function are ellipses, like for the multivariate normal,
- ... which is a particular case with $g(u) = e^{-u/2}$.
- The covariance matrix (if finite) is proportional to $\boldsymbol{\Sigma}$ (for this reason, the correlations are $\sigma_{i,j}/(\sigma_{i,i}\sigma_{j,j})$ with $\sigma_{i,j}$ the elements of Σ).

Elliptical bivariate $t(5)$ & normal (same covariance matrix)





... and the level curves



Factor models

Consider $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{H}\boldsymbol{Z}$ beyond $\dim \boldsymbol{X} = \dim \boldsymbol{Z} = n$.

In particular, the case $\dim \boldsymbol{Z} = r < n$ may be interesting:

- Of course, $\mathbf{H}$ (and thus $\boldsymbol{\Sigma}$) is of rank $r < n$.
- To alleviate this, add some randomness, say

  $$\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{H}\boldsymbol{Z} + \boldsymbol{E}$$

  where $\boldsymbol{E}$ is a vector of uncorrelated RVs, also uncorrelated with $\boldsymbol{Z}$.
- This implies

  $$\mathrm{Cov}(\boldsymbol{X}) = \mathbf{H}\,\mathrm{Cov}(\boldsymbol{Z})\mathbf{H}' + \mathrm{Cov}(\boldsymbol{E})$$

  where $\mathrm{Cov}(\boldsymbol{E})$ is diagonal, making $\mathrm{Cov}(\boldsymbol{X})$ full-rank again.

Such models are flexible (imagine e.g. a large set of variables depending on a small number of "**common factors**") – and work for discrete RVs too.

As opposed to multivariate mixtures, here we combine random variables linearly and not densities.

Towards an encompassing class

Take the univariate normal density, $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ and rewrite it as

$$f(x) = \exp\left(\frac{\mu}{\sigma^2} \cdot x - \frac{1}{2\sigma^2} \cdot x^2 - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{\mu^2}{\sigma^2}\right)$$

We may do the same for other distributions:

- Bernoulli: $f(x) = \exp\left(x \log \frac{p}{1-p} + \log p + \log \mathbb{I}_{\{0,1\}}(x)\right)$
- Poisson: $f(x) = \exp\left(x \log\lambda - \lambda - \log x! + \log \mathbb{I}_{\mathbb{N}}(x)\right)$
- Exponential:[1] $f(x) = \exp\left(-x\lambda + \log\lambda + \log \mathbb{I}_{\mathbb{R}_+}(x)\right)$

... and note how parameters and density arguments interact.

[1] In the λ and not the θ parameterization, $f(x) = \lambda \exp(-\lambda x)$.

The exponential class

Definition

The pdf $f(\boldsymbol{x};\boldsymbol{\theta})$ is a member of the exponential class of pdfs iff it has the form

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \begin{cases} \exp\left\{\boldsymbol{\Sigma}_{i=1}^{k} c_i\left(\boldsymbol{\theta}\right) g_i(\boldsymbol{x}) + d\left(\boldsymbol{\theta}\right) + z(\boldsymbol{x})\right\} & \text{for } \boldsymbol{x} \in A \\ 0 \quad \text{otherwise} \end{cases},$$

where

$$\begin{array}{rcl} \boldsymbol{x} & = & (x_1, \ldots, x_n)'; \\ \boldsymbol{\theta} & = & (\theta_1, \ldots, \theta_k)'; \\ c_i(\boldsymbol{\theta}),\ d(\boldsymbol{\theta}) & : & \text{real-valued functions of } \boldsymbol{\theta} \text{ that do not depend on } \boldsymbol{x}; \\ g_i(\boldsymbol{x}),\ z(\boldsymbol{x}) & : & \text{real-valued functions of } \boldsymbol{x} \text{ that do not depend on } \boldsymbol{\theta}; \\ A \subset \mathbb{R}^n & : & \text{a range/support which } \textbf{does not depend} \text{ on } \boldsymbol{\theta}. \end{array}$$

Members of the exponential class

For $\mathcal{N}(\mu, \sigma^2)$ with $n = 1$ and $k = 2$ ($\#$ of parameters),

$$c_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}, \qquad c_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}, \qquad g_1(x) = x, \qquad g_2(x) = x^2;$$

$$d(\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{\mu^2}{\sigma^2}, \qquad z(x) = 0, \qquad A = \mathbb{R}.$$

The multivariate normal also fits the exponential class, btw. And so do
Bernoulli, binomial, multinomial, negative binomial, Poisson, geometric, gamma, chi-square, exponential, beta, etc.

Distributions that do not belong to the exponential class are, e.g.:
discrete uniform, continuous uniform, hypergeometric.

The exponential class of densities is a very popular model (and also has nice inferential properties; see Advanced Statistics II).

Outline

Modelling dependence

Take the random vector $(Y, \boldsymbol{X}')'$ with joint density $f_{Y,\boldsymbol{X}}$.

- If Y and $\boldsymbol{X}$ are statistically independent,

$$f_{Y,\boldsymbol{X}} = f_Y f_{\boldsymbol{X}}$$

  where $f_Y, f_{\boldsymbol{X}}$ are the respective marginal distributions.
- If not, we may write

$$f_{Y,\boldsymbol{X}} = f_{Y|\boldsymbol{X}} f_{\boldsymbol{X}},$$

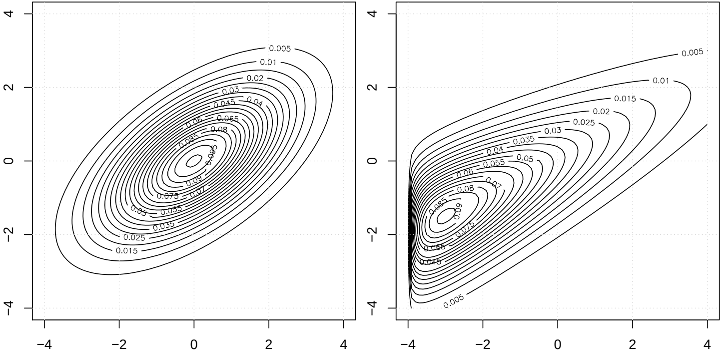
  where $f_{Y|\boldsymbol{X}}$ is the conditional distribution of Y given $\boldsymbol{X}$.
- The conditional distribution changes, in general, with $\boldsymbol{X}$!

This helps understand how (features of) Y depend(s) on $\boldsymbol{X}$, and use such knowledge to set up models suitable for dependent data.[2]

[2] Beware of causal interpretations, though!

Joint distributions

Different marginal distributions imply different joint ones.



But: If only dependence is of interest, f_X need not receive any attention.[3]

[3] Copulas offer another way of decomposing joint distributions; see the lecture notes.

The linear conditional normal model

For the multivariate normal,

- the conditional distribution is normal, and
- the conditional expectation is linear.

We may write

$$Y = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} + E$$

where $E \sim \mathcal{N}(0, \sigma^2)$, independent of $\boldsymbol{X}$.

We then obtain the linear conditionally normal model

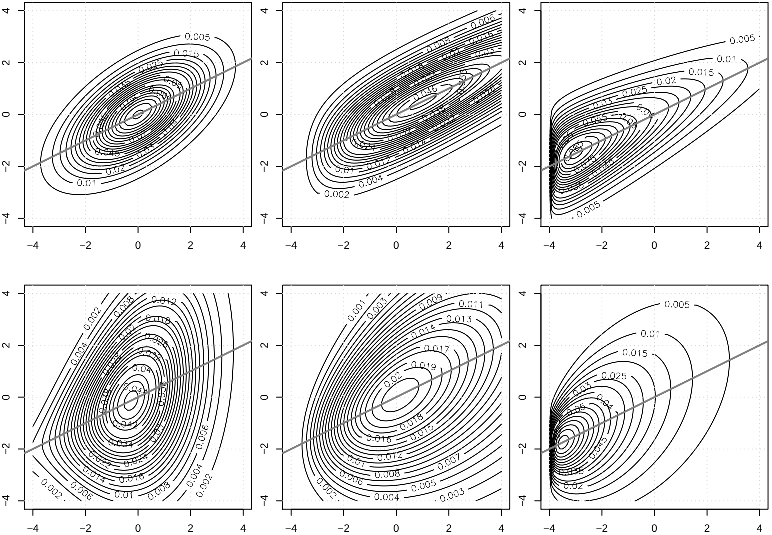
$$Y|\boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}, \sigma^2).$$

(Y is only marginally normal if $\boldsymbol{X}$ is normal, independent of the errors.)

For non-Gaussian E or varying σ^2, we in fact *only* model $\mathrm{E}(Y|\boldsymbol{X})$![4]

[4] This is what people usually understand under a **regression model**.

Same linear regression curve, different cond. distributions



Possible extensions

Such models for the conditional mean can be used

- for forecasting when outcomes for $\boldsymbol{X}$ are observed, and
- also for causal analyses (with some care of course; see any econometrics course).

This makes them quite useful...

We would also like to have such models for

- other conditional functionals, say quantiles,
- other distributions than the normal, or
- other functional forms than linearity.

Each relaxation (not to mention all at the same time) brings up interesting models...

Conditional quantiles

Let Y and $\boldsymbol{X}$ have a continuous joint distribution.

- Then, q_p such that $\mathrm{P}\left(Y \leq q_p\right) = p$ is the marginal p-quantile of Y.
- When focussing on conditional distributions, we naturally obtain conditional quantiles...

So we call $q_p(\boldsymbol{x})$ the conditional p-quantile of Y given $\boldsymbol{X} = \boldsymbol{x}$ if

$$\mathrm{P}\left(Y \leq q_p(\boldsymbol{x}) | \boldsymbol{X} = \boldsymbol{x}\right) = p.$$

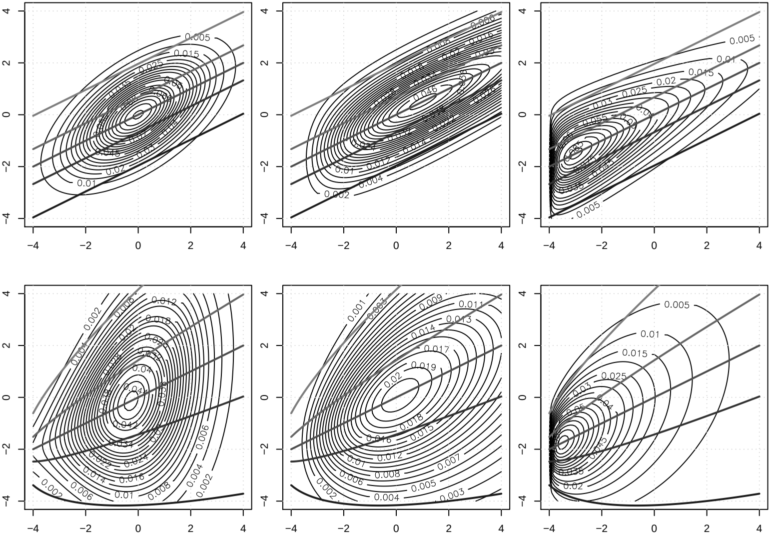
If q_p is linear in $\boldsymbol{x}$, we obtain a linear quantile regression model.
Note that q_p will be different for different levels p!

We may rewrite this as

$$Y = q_p(\boldsymbol{X}) + E_p$$

where E_p has zero conditional quantile given $\boldsymbol{X}$.

Quantile functions for $p \in \{0.025, 0.25, 0.5, 0.75, 0.975\}$



Outline

Other types of data

A linear regression model is clearly a bad model for binary or count data (or for durations for that matter).

In the univariate case, we used e.g. the Bernoulli or Gamma distributions.

- So we would have to set up models where the conditional distribution is Bernoulli etc.
- Like for the normal regression, we make parameters of these distributions depend on $\boldsymbol{X}$!

We'll do this in the framework of the exponential class of densities, which leads us to **generalized linear models**.

The GLM

To keep things nicely interpretable, we'll resort to a special version of the univariate exponential class. Start with

$$f(y,\theta) = \exp\left(\theta \cdot y - b(\theta) + z(y)\right).$$

(We say the density is in **canonical form** iff $c(\theta) = \theta$.)

You can represent the Bernoulli and Poisson distributions this way, but not the normal. So add a second parameter to obtain a so-called overdispersed version, namely

$$f(y,\theta,\psi) = \exp\left(\frac{\theta \cdot y - b(\theta)}{a(\psi)} + d(y,\psi) + z(y)\right).$$

The Bernoulli case

If Y is Bernoulli distributed,

$$f(y) = \exp\left(y \log\left(\frac{p}{1-p}\right) + \log(1-p) + \log \mathbb{I}_{\{0,1\}}(y)\right).$$

With $\theta := \log\left(\frac{p}{1-p}\right)$, we have an exponential family with no extra parameter ψ,

$$a(\psi) = 1, \quad b(\theta) = -\log(1-p) = \log(1+e^{\theta}), \quad d(y,\psi) = 0.$$

(We call $\log\left(\frac{p}{1-p}\right)$ the logit or log-odds transform, and its inverse $e^{\theta}/(1+e^{\theta})$ the logistic transform.)

... and the rest

Thus,

- Poisson:

$$\theta = \log\lambda,\ b(\theta) = \lambda = e^{\theta},\ b'(\theta) = e^{\theta},\ a(\psi) = 1;$$

- Exponential:[5]

$$\theta = \lambda,\ b(\theta) = \log\lambda = \log\theta,\ b'(\theta) = \frac{1}{\theta},\ a(\psi) = -1:$$

- Gaussian:

$$\theta = \mu,\ b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2},\ b'(\theta) = \theta,\ a(\psi) = \sigma^2.$$

[5] Again, in λ parameterization to avoid confusions.

Properties of the canonical form

Lemma

Regularity conditions assumed, we have for the above representation

$$\begin{aligned} \mathrm{E}(Y) &= b'(\theta), \\ \mathrm{Var}(Y) &= b''(\theta)a(\psi). \end{aligned}$$

For linear regression we had

$$\mathrm{E}(Y|\boldsymbol{X}) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}.$$

Putting this in GLM language, we have equivalently

$$b'(\theta) := \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} = \mu(\boldsymbol{X})$$

This way obtain a conditional model for Y given $\boldsymbol{X} = \boldsymbol{x}$.

The link function

What we did for the Gaussian was to choose θ to be a linear function of $\boldsymbol{X}$.

Generally we set $\mathrm{E}(Y|\boldsymbol{X}) = b'(\theta) = G(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X})$ where G is called **link function**.

Then, $b'(\theta)$ gives **the canonical** link:

- Gaussian regression: G is the identity function;
- (Exponential) hazard model: G is the reciprocal function;
- Poisson regression: G is the exponential link;
- (Bernoulli) Logit regression: G is the logistic transform $e^{\theta}/(1 + e^{\theta})$

One may even use other link functions instead of the above canonical ones.

And of course linearity of b' in $\boldsymbol{X}$ may be relaxed.

More nonlinearity

Sofar, the GLM used a transformation of a linear combination of the $\boldsymbol{X}$s[6] to model conditional means.

But there is nothing that keeps us from making more generalizations:

- We may take nonlinear functions of $\boldsymbol{X}$
  - E.g. for the Gaussian GLM, this leads to nonlinear regression models
- We may also let ψ also depend on $\boldsymbol{X}$
  - This usually amounts modeling the conditional mean *and* the conditional variance
  - ... leading for the Gaussian GLM to location-scale models

Conditional location-scale models can also be quite useful in practice...

[6] Such models are called single-index models, btw.

Conditional location-scale models

Write

$$Y = m(\boldsymbol{X}) + \sigma(\boldsymbol{X})E$$

where E has zero conditional mean and unit conditional variance.[7] Then,

$$\mathrm{E}(Y|\boldsymbol{X}) = m(\boldsymbol{X}), \quad \mathrm{Var}(Y|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}).$$

If E and $\boldsymbol{X}$ are independent, there's a direct implication for quantiles:

- Let $q_{p,E}$ be the p-quantile of E: under independence of E and $\boldsymbol{X}$,

$$\mathrm{P}(E - q_{p,E} < 0) = p = \mathrm{P}(E - q_{p,E} < 0|\boldsymbol{X})$$

- This implies $\mathrm{P}(Y < \mu(\boldsymbol{x}) + q_{p,E}\sigma(\boldsymbol{x})|\boldsymbol{X} = \boldsymbol{x}) = p$, or

$$Y = \mu(\boldsymbol{x}) + q_{p,E}\sigma(\boldsymbol{x}) + U_p$$

  where $U_p = \sigma(\boldsymbol{x})(E - q_{p,E})$ has zero conditional p-quantile.

[7] This is satisfied if E is standardized and independent of $\boldsymbol{X}$.

Outline

Coming up

Basic asymptotics