Notes on Robust Student t Regression

Max Leung

April 14, 2024

1 Arbitrary Elliptically Symmetric Family of Densities

Suppress notation i for observation i. $dim(\boldsymbol{y}) = k$

1.1 Probability Density Function

$$p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu) = |\boldsymbol{\Omega}|^{-1/2} g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)$$

1.1.1 Special Case: k-variate generalized t distribution

$$g(s,\nu) = \frac{\Gamma((\nu+k)/2)}{\Gamma(1/2)^k \Gamma(\nu/2)\nu^{k/2}}(1+\frac{s}{\nu})^{-(\nu+k)/2}$$

$$\boldsymbol{y} \sim t_k(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Omega}, \nu)$$

Property: $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{\mu}(\boldsymbol{\theta})$

1.1.2 Special Case: k-variate generalized power-exponential family

$$g(s,\nu) = c(\nu)e^{-s^{\nu}/2}$$

1.2 Gradient Vector

Assume $\boldsymbol{\mu}$ is a function of $\boldsymbol{\theta}$ i.e., $\boldsymbol{\mu}(\boldsymbol{\theta})$

$$\begin{aligned}
\frac{\partial ln[p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu)]}{\partial \boldsymbol{\theta}} &= \frac{\partial ln[|\boldsymbol{\Omega}|^{-1/2} g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)]}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial (ln[|\boldsymbol{\Omega}|^{-1/2}] + ln[g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)])}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial ln[g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)]}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial ln[g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)]}{\partial g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)} \frac{\partial g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{g}(\frac{\partial g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)}{\partial (\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})} \frac{\partial (\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{\partial \boldsymbol{\theta}} + \frac{\partial g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)}{\partial \nu} \frac{\partial \nu}{\partial \boldsymbol{\theta}}) \\
&= \frac{1}{g}(g_1 \cdot \frac{\partial (\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{\partial \boldsymbol{\theta}} + \frac{\partial g((\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \nu)}{\partial \nu} \cdot 0) \\
&= \frac{g_1}{g} \frac{\partial (\boldsymbol{y}-\boldsymbol{\mu})'}{\partial \boldsymbol{\theta}} (\boldsymbol{\Omega}^{-1} + \boldsymbol{\Omega}^{-1'})(\boldsymbol{y}-\boldsymbol{\mu}) \\
&= \frac{g_1}{g}(-\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}})(2 \cdot \boldsymbol{\Omega}^{-1})(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\theta})) \\
&= -2\frac{g_1}{g} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}} \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))
\end{aligned}$$

1.2.1 Special Case: k-variate generalized t distribution

$$
\begin{aligned}
g_1 &= \frac{d\frac{\Gamma((\nu+k)/2)}{\Gamma(1/2)^k\Gamma(\nu/2)\nu^{k/2}}(1+\frac{s}{\nu})^{-(\nu+k)/2}}{ds} \\
&= -\frac{\Gamma((\nu+k)/2)}{\Gamma(1/2)^k\Gamma(\nu/2)\nu^{k/2}}((\nu+k)/2)(1+\frac{s}{\nu})^{-(\nu+k)/2-1}\frac{1}{\nu}
\end{aligned}
$$

$$
\begin{aligned}
\frac{g_1}{g} &= \frac{-\frac{\Gamma((\nu+k)/2)}{\Gamma(1/2)^k\Gamma(\nu/2)\nu^{k/2}}((\nu+k)/2)(1+\frac{s}{\nu})^{-(\nu+k)/2-1}\frac{1}{\nu}}{\frac{\Gamma((\nu+k)/2)}{\Gamma(1/2)^k\Gamma(\nu/2)\nu^{k/2}}(1+\frac{s}{\nu})^{-(\nu+k)/2}} \\
&= -((\nu+k)/2)(1+\frac{s}{\nu})^{-1}\frac{1}{\nu} \\
&= -\frac{1}{2}(\nu+k)\frac{\nu}{\nu+s}\frac{1}{\nu} \\
&= -\frac{1}{2}\frac{\nu+k}{\nu+s}
\end{aligned}
$$

where $s = (\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$

1.3 Maximum Likelihood Estimation

Assume independent $\boldsymbol{y}_i$ with common density p

$$
\begin{aligned}
\frac{\partial lnL}{\partial\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}_{mle}} = \frac{\partial ln\prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{\mu}_i,\boldsymbol{\Omega}_i,\nu)}{\partial\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}_{mle}} &= \mathbf{0} \\
\sum_{i=1}^N \frac{\partial ln[p(\boldsymbol{y}_i|\boldsymbol{\mu}_i,\boldsymbol{\Omega}_i,\nu)]}{\partial\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}_{mle}} &= \mathbf{0} \\
\sum_{i=1}^N \frac{g_{1,i}}{g_i}\frac{\partial\boldsymbol{\mu}_i(\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}_{mle}}\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}}_{mle})) &= \mathbf{0}
\end{aligned}
$$

1.3.1 Special Case: k-variate generalized t distribution

$s_i = (\boldsymbol{y}_i-\boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}}_{mle}))'\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}}_{mle}))$

$$
\begin{aligned}
\sum_{i=1}^N(-\frac{1}{2}\frac{\nu+k_i}{\nu+s_i})\frac{\partial\boldsymbol{\mu}_i(\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}_{mle}}\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}}_{mle})) &= \mathbf{0} \\
\sum_{i=1}^N\frac{\nu+k_i}{\nu+s_i}\frac{\partial\boldsymbol{\mu}_i(\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}_{mle}}\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}}_{mle})) &= \mathbf{0}
\end{aligned}
$$

The resulting ML estimator is robust in the sense that outlier observation with large s_i has smaller weight in the log likelihood function.

If $k_i = 1$ for $\forall i$, $y_i \sim t_1(\mu(\boldsymbol{\theta},\boldsymbol{x}_i),\sigma_i^2,\nu)$ where $\mu(\boldsymbol{\theta},\boldsymbol{x}_i)$ can be linear $\boldsymbol{x}_i'\boldsymbol{\theta}$

$$
\sum_{i=1}^N\frac{\nu+1}{\nu+s_i}\frac{\partial\mu(\boldsymbol{\theta},\boldsymbol{x}_i)'}{\partial\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}_{mle}}\sigma_i^{-2}(y_i-\mu(\widehat{\boldsymbol{\theta}}_{mle},\boldsymbol{x}_i)) = 0
$$

where $s_i = (y_i-\mu(\widehat{\boldsymbol{\theta}}_{mle},\boldsymbol{x}_i))^2/\sigma_i^2$

The following R code shows an example of ML estimation of robust student t regression.

```
library(tidyverse)

t_reg_mle <- function (
    y_,
    X_,
    init_par_,
    sigma_ = NULL,
    intercept_ = TRUE,
```

```
    method_ = "BFGS"
  ) {

  if (!is.numeric(y_)) stop("y_ should be a numeric vector")
  if (!is.matrix(X_)) stop("X_ should be a matrix")
  if (intercept_) X_ <- cbind(1, X_)
  if (!is.numeric(init_par_)) stop("init_par_ should be a numeric vector")
  if (is.null(sigma_) && (length(init_par_) != ncol(X_) + 2)) stop("wrong init_par_")
  if (!is.null(sigma_) && (length(init_par_) != ncol(X_) + 1)) stop("wrong init_par_")

  negative_lnL <- function (par, y, X) {
    beta <- par[1:ncol(X)]
    df <- par[ncol(X) + 1]

    if (is.null(sigma_)) {
      sigma <- par[ncol(X) + 2]
    } else if (!is.na(sigma_) && is.numeric(sigma_) && sigma_ > 0) {
      sigma <- sigma_
    } else {
      stop("sigma_ should be positve real number")
    }

    -sum(log(dt((y - X %*% beta) / sigma^2, df = df) / sigma^2))
  }

  optim(
    par = init_par_,
    fn = negative_lnL,
    y = y_, X = X_,
    method = method_,
    hessian = TRUE
  )
}

X_matrix <-
  d %>%
  select(x1, x2, x3) %>%
  as.matrix()

y_vector <- d[["y"]]

mle <- t_reg_mle(y_ = y_vector, X_ = X_matrix, init_par_ = c(0.3, 0.2, -0.4, 0.5, 3, 1))
mle$par
mle$hessian %>% solve() %>% diag() %>% sqrt()
```

1.4 Student-t is a mix density

The density of k-variate generalized t random variable can also be written as a mixture of multivariate normal and scaled inverse chi-square densities.

$$
\begin{aligned}
t(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu) &= \int_0^\infty N(\boldsymbol{y}|\boldsymbol{\mu}, z \cdot \boldsymbol{\Omega})\chi^{-2}(z|\nu, 1)dz \\
&= \int_0^\infty p(\boldsymbol{y}, z|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu)dz
\end{aligned}
$$

Note that $IG(z|\frac{\nu}{2}, \frac{\nu}{2}) = \chi^{-2}(z|\nu, 1)$ where IG = Inverse Gamma. $IG(z|a, b) = \frac{b^a}{\Gamma(a)}z^{-(a+1)}exp(-b/z)$

If $k = 1$,

$$
\begin{aligned}
t(y|\mu, \sigma^2, \nu) &= \int_0^\infty N(y|\mu, z \cdot \sigma^2)\chi^{-2}(z|\nu, 1)dz \\
&= \int_0^\infty p(y, z|\mu, \sigma^2, \nu)dz
\end{aligned}
$$

2 EM estimation of robust regression with univariate generalized t distribution

Regarding z_i as missing data, assume ν is known (or specify a grid of possible ν values).

2.1 E Step

$$
\begin{aligned}
ln[p(y_i, z_i|\overbrace{\mu(\boldsymbol{\theta}, \boldsymbol{x}_i)}^{\boldsymbol{\theta},\boldsymbol{x}_i}, \sigma_i^2, \nu)] &= ln[N(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i, z_i \cdot \sigma_i^2)\chi^{-2}(z_i|\nu, 1)] \\
&= ln[N(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i, z_i \cdot \sigma_i^2)] + ln[\chi^{-2}(z_i|\nu, 1)] \\
&= ln[\frac{1}{\sqrt{z_i \cdot \sigma_i^2 \cdot 2\pi}} exp(-\frac{1}{2}\frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{z_i\sigma_i^2})] + ln[\frac{1}{\Gamma(\nu/2)}(\frac{\nu}{2})^{\frac{\nu}{2}} z_i^{-\nu/2-1} exp(-\frac{\nu}{2z_i})] \\
&= -\frac{1}{2}ln(z_i \cdot \sigma_i^2 \cdot 2\pi) - \frac{1}{2}\frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{z_i\sigma_i^2} + (-\frac{\nu}{2} - 1)ln[z_i]^{-\nu/2-2} - \frac{\nu}{2z_i} + constant
\end{aligned}
$$

Thus, given independence

$$
\begin{aligned}
\mathbb{E}[ln[p(\boldsymbol{y}, z_i|\boldsymbol{\theta}, \boldsymbol{x}_i, \sigma_i^2, \nu)]|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu] &= \mathbb{E}[ln[\prod_{i=1}^N p(y_i, z_i|\boldsymbol{\theta}, \boldsymbol{x}_i, \sigma_i^2, \nu)]|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu] \\
&= \mathbb{E}[\sum_{i=1}^N ln[p(y_i, z_i|\boldsymbol{\theta}, \boldsymbol{x}_i, \sigma_i^2, \nu)]|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu] \\
&= \mathbb{E}[\sum_{i=1}^N\{-\frac{1}{2}ln(z_i \cdot \sigma_i^2 \cdot 2\pi) - \frac{1}{2}\frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{z_i\sigma_i^2} + (-\frac{\nu}{2} - 1)ln[z_i]^{-\nu/2-2} - \frac{\nu}{2z_i} + constant\}| \\
&\quad y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu]
\end{aligned}
$$

Omitting terms without $\boldsymbol{\theta}$, as those will be differentiated away in M step

$$
\begin{aligned}
&\approx \mathbb{E}[-\frac{1}{2}\frac{\sum_{i=1}^N(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{z_i\sigma_i^2}|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu] \\
&= -\sum_{i=1}^N \underbrace{\mathbb{E}[\frac{1}{z_i}|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu]}_{w_i} \frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{2\sigma_i^2}
\end{aligned}
$$

In order to evaluate w_i, we have to find out the posterior distribution $z_i|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu$

$$
\begin{aligned}
p(z_i|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu) &\propto p(y_i|z_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2)p(z_i|\nu) \\
&= N(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i, z_i \cdot \sigma_i^2)\chi^{-2}(z_i|\nu, 1) \\
&= \frac{1}{\sqrt{z_i \cdot \sigma_i^2 \cdot 2\pi}} exp(-\frac{1}{2}\frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{z_i\sigma_i^2}) \cdot \frac{1}{\Gamma(\nu/2)}(\frac{\nu}{2})^{\frac{\nu}{2}} z_i^{-\nu/2-1} exp(-\frac{\nu}{2z_i}) \\
&= z_i^{-1/2} z_i^{-\nu/2-1} exp(-\frac{1}{2}\frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{z_i\sigma_i^2}) exp(-\frac{\nu}{2z_i}) \cdot C \\
&= z_i^{-(v+1)/2-1} exp(-\frac{1}{2z_i}(\frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{\sigma_i^2} + \nu)) \cdot C \\
&= z_i^{-[(v+1)/2+1]} exp(-[\frac{1}{2}(\frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{\sigma_i^2} + \nu)]/z_i) \cdot C
\end{aligned}
$$

Given the fact that inverse gamma is the conjugate prior of normal-inverse-gamma mix, $z_i|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu \sim IG(\frac{\nu+1}{2}, \frac{(y_i-\mu(\boldsymbol{\theta},\boldsymbol{x}_i))^2/\sigma_i^2+\nu}{2})$. Therefore, $\frac{1}{z_i} \sim G(\frac{\nu+1}{2}, \frac{(y_i-\mu(\boldsymbol{\theta},\boldsymbol{x}_i))^2/\sigma_i^2+\nu}{2})$. By the property of Gamma random variable, $w_i = \mathbb{E}[\frac{1}{z_i}|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \sigma_i^2, \nu] = (\frac{\nu+1}{2})/(\frac{(y_i-\mu(\boldsymbol{\theta},\boldsymbol{x}_i))^2/\sigma_i^2+\nu}{2}) = \frac{\nu+1}{(y_i-\mu(\boldsymbol{\theta},\boldsymbol{x}_i))^2/\sigma_i^2+\nu}$

2.2 M Step

$$
\begin{aligned}
\boldsymbol{\theta}^{t+1} &= arg\ max_{\boldsymbol{\theta}} \mathbb{E}[ln[p(\boldsymbol{y}, z_i|\boldsymbol{\theta}, \boldsymbol{x}_i, \sigma_i^2, \nu)]|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}^t, \sigma_i^{2,t}, \nu] \\
&= arg\ max_{\boldsymbol{\theta}}\{-\sum_{i=1}^N w_i^t \frac{(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{2\sigma_i^{2,t}}\} \\
&= arg\ min_{\boldsymbol{\theta}}\{\sum_{i=1}^N \frac{w_i^t}{\sigma_i^{2,t}}(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2\}
\end{aligned}
$$

where $w_i^t = \frac{\nu+1}{(y_i-\mu(\boldsymbol{\theta}^t,\boldsymbol{x}_i))^2/\sigma_i^{2,t}+\nu}$ and $\sigma_i^{2,t+1} = \sum_{i=1}^N \frac{w_i^t}{\sigma_i^{2,t}}(y_i - \mu(\boldsymbol{\theta}^{t+1}, \boldsymbol{x}_i))^2/N$

It is a weighted non-linear least square (NLS) problem with weight $\frac{w_i}{\sigma_i^2}$. Thus, some says that EM algorithm is Iteratively Re-weighted NLS. Its First Order Condition (FOC) is similar to MLE's FOC.

$$
\begin{aligned}
\frac{\partial \sum_{i=1}^N \frac{w_i^t}{\sigma_i^{2,t}}(y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^{t+1}} &= \mathbf{0} \\
\sum_{i=1}^N \frac{w_i^t}{\sigma_i^{2,t}} \frac{\partial (y_i - \mu(\boldsymbol{\theta}, \boldsymbol{x}_i))^2}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^{t+1}} &= \mathbf{0} \\
\sum_{i=1}^N \frac{w_i^t}{\sigma_i^{2,t}} \cdot (-2) \cdot (y_i - \mu(\boldsymbol{\theta}^{t+1}, \boldsymbol{x}_i)) \frac{\partial \mu(\boldsymbol{\theta}, \boldsymbol{x}_i)'}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^{t+1}} &= \mathbf{0} \\
\sum_{i=1}^N w_i^t \frac{\partial \mu(\boldsymbol{\theta}, \boldsymbol{x}_i)'}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^{t+1}} \sigma_i^{-2,t}(y_i - \mu(\boldsymbol{\theta}^{t+1}, \boldsymbol{x}_i)) &= \mathbf{0}
\end{aligned}
$$

3 Reference

Lange, K. L., Roderick J. A. Little, & Jeremy M. G. Taylor. (1989). Robust Statistical Modeling Using the t Distribution. Journal of the American Statistical Association, 84(408), 881–896. https://doi.org/10.2307/2290063

Kevin P. Murphy. (2023). Probabilistic Machine Learning: Advanced Topics, MIT Press