Student's $\mathcal{T}$ Distribution

Prof. Nicholas Zabaras
School of Engineering
University of Warwick
Coventry CV4 7AL
United Kingdom

Email: nzabaras@gmail.com
URL: http://www.zabaras.com/

August 7, 2014

Contents

- Following closely Chris Bishops' PRML book, Chapter 2
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapter 2

Gamma as a Conjugate Prior

- We have seen that the conjugate prior for the precision of a Gaussian is given by a *Gamma* distribution. If we have a univariate Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ together with a prior *Gamma*$(\tau\,|a, b)$ and we integrate out the precision, we obtain the marginal distribution of x

$$p(x|\mu,a,b)=\int_0^\infty \mathcal{N}\left(x|\mu,\tau^{-1}\right)\mathcal{Gamma}\left(\tau|a,b\right)d\tau =$$

$$=\int_0^\infty \left(\frac{\tau}{2\pi}\right)^{1/2}\exp\left(-\frac{\tau}{2}\left(x-\mu\right)^2\right)\frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau}d\tau =$$

- Introduce the transformation $z=\underbrace{\left[b+\frac{1}{2}\left(x-\mu\right)^2\right]}_{A}\tau$ to simplify as:

$$p(x|\mu,a,b)=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\int_0^\infty \tau^{1/2}\exp\left(-z\right)\tau^{a-1}d\tau =$$

$$=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\frac{1}{A^{1/2+a-1+1}}\int_0^\infty z^{1/2}\exp\left(-z\right)z^{a-1}dz$$

Student's $\mathcal{T}$ Distribution

$$p(x\,|\,\mu,a,b) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\frac{1}{A^{1/2+a-1+1}}\int_0^\infty z^{1/2}\exp(-z)\,z^{a-1}dz$$

$$= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b+\frac{1}{2}(x-\mu)^2\right]^{-a-1/2}\int_0^\infty \exp(-z)\,z^{a-1/2}dz$$

- Recalling the definition of the Gamma function: $\Gamma(a) = \int_0^\infty \exp(-z)\,z^{a-1}dz$

$$p(x\,|\,\mu,a,b) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b+\frac{1}{2}(x-\mu)^2\right]^{-a-1/2}\Gamma(a+\frac{1}{2})$$

- It is common to redefine the parameters in this distribution as: $\upsilon = 2a,\ \lambda = \dfrac{a}{b}$

$$p(x\,|\,\mu,\lambda,\upsilon) = \frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\upsilon}\right]^{-\upsilon/2-1/2}$$

Student's $\mathcal{T}$ Distribution

$$p(x\,|\,\mu,\lambda,\upsilon)=\frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]^{-\upsilon/2-1/2}$$

- The parameter λ is called the precision of the t-distribution, even though it is not in general equal to the inverse of the variance (see below on behavior as $v \rightarrow \infty$).

- The parameter v is called the degrees of freedom.

- For the particular case of v = 1, the $\mathcal{T}$-distribution reduces to the Cauchy distribution.

- In the limit $v \rightarrow \infty$, the t-distribution $\mathcal{T}(x|\mu, \lambda, v)$ becomes a Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$ with mean μ and precision λ.

For $\nu \to \infty$, $\mathcal{T}(x|\mu, \lambda, \nu)$ Becomes a Gaussian

$$p(x|\mu,\lambda,\upsilon)=\frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]^{-\upsilon/2-1/2}$$

- We first write the distribution as follows:

$$\mathcal{T}(x|\mu,\lambda,\upsilon)\propto\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]^{-\upsilon/2-1/2}=\exp\left\{-\frac{\upsilon+1}{2}\ln\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]\right\}$$

- For large ν we can approximate the log as follows:

$$\mathcal{T}(x|\mu,\lambda,\upsilon)\propto\exp\left\{-\frac{\upsilon+1}{2}\left[\frac{\lambda\left(x-\mu\right)^2}{\upsilon}+O(\upsilon^{-2})\right]\right\}=\exp\left\{-\frac{\lambda\left(x-\mu\right)^2}{2}+O(\upsilon^{-1})\right\}$$

- In the limit $\nu \to \infty$, the $\mathcal{T}$-distribution $\mathcal{T}(x|\mu, \lambda, \nu)$ is indeed a Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$ with mean μ and precision λ. The normalization of the $\mathcal{T}$ is valid in this limit as well (so the Gaussian obtained is normalized).

Student's $\mathcal{T}$ Distribution

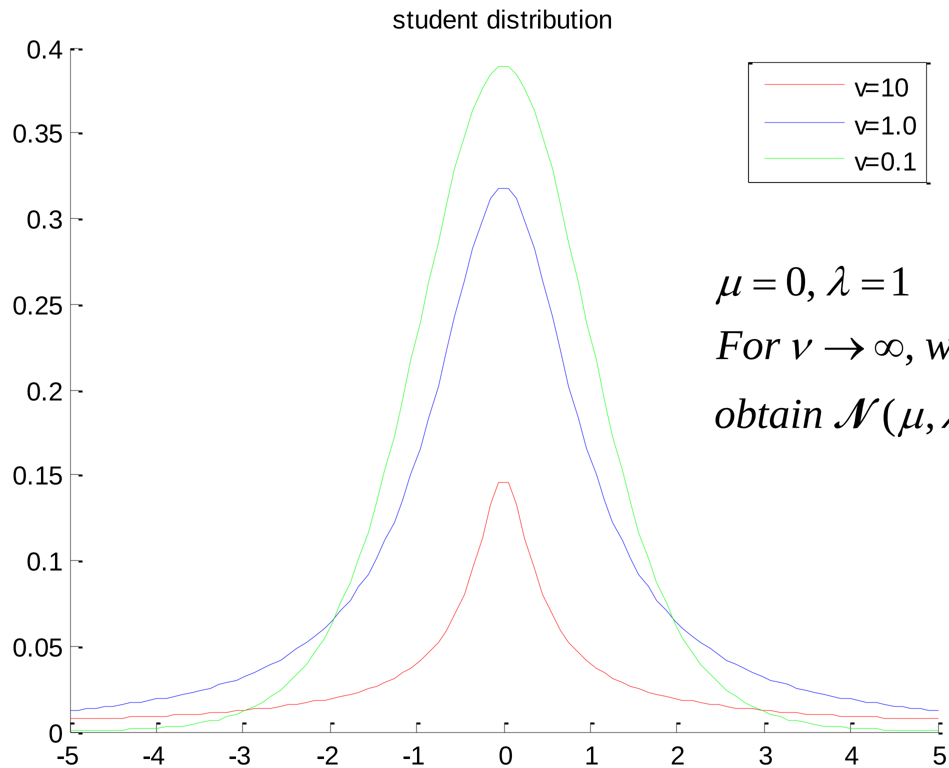
$$p(x|\mu,\lambda,\upsilon)=\frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda\left(x-\mu\right)^{2}}{\upsilon}\right]^{-\upsilon/2-1/2}$$

$Mean: \mu, \upsilon > 1$

$Mode: \mu$

$$Var: \frac{\upsilon\sigma^2}{\upsilon-2} = \frac{\upsilon}{\lambda\left(\upsilon-2\right)}, \upsilon > 2$$

$$\lambda = \sigma^{-2}$$



$\mu = 0, \lambda = 1$

$For\ \nu \to \infty,\ we$

$obtain\ \mathcal{N}(\mu, \lambda^{-1})$

MatLab Code

Student's $\mathcal{T}$ Vs the Gaussian

- We plot:

$$\mathcal{N}\left(x\,|\,0,1\right), \mathcal{T}(x\,|\,0,1,1), \mathcal{Lap}\left(x\,|\,0,1/\sqrt{2}\right)$$

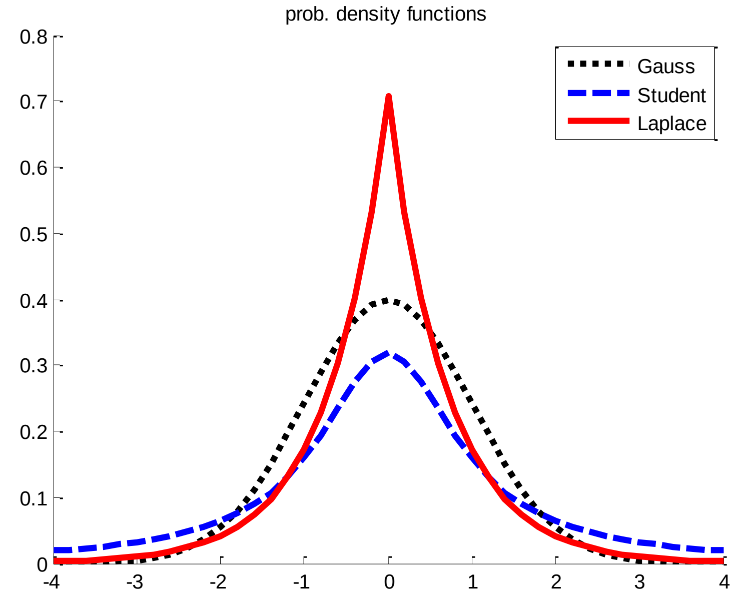
- *The mean and variance of the Student's is undefined for ν=1*.

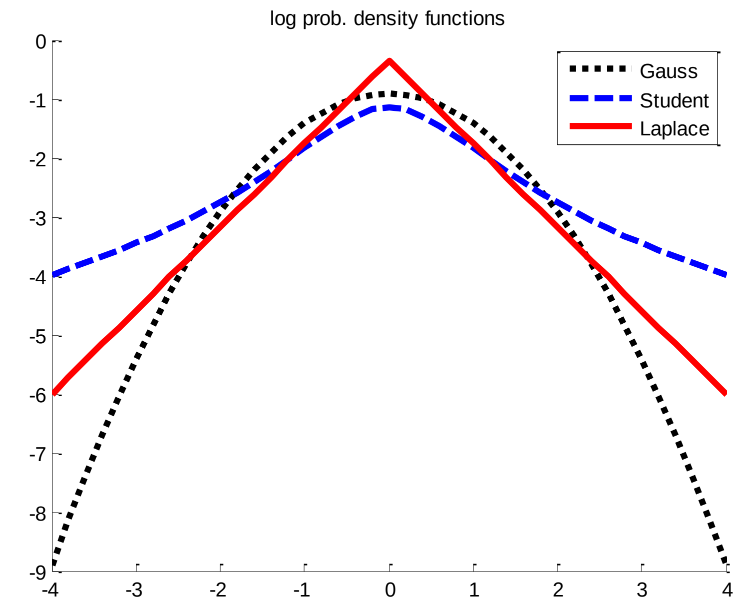
- Logs of the PDFs. *The Student's is NOT log concave.*

  Run MatLab function *studentLaplacePdfPlot* from Kevin Murphys' PMTK

- *When ν=1, the distribution is known as Cauchy or Lorentz. Due to its heavy tails, the mean does not converge.*

- ***Recommended to use ν=4.***





Student's $\mathcal{T}$ Distribution

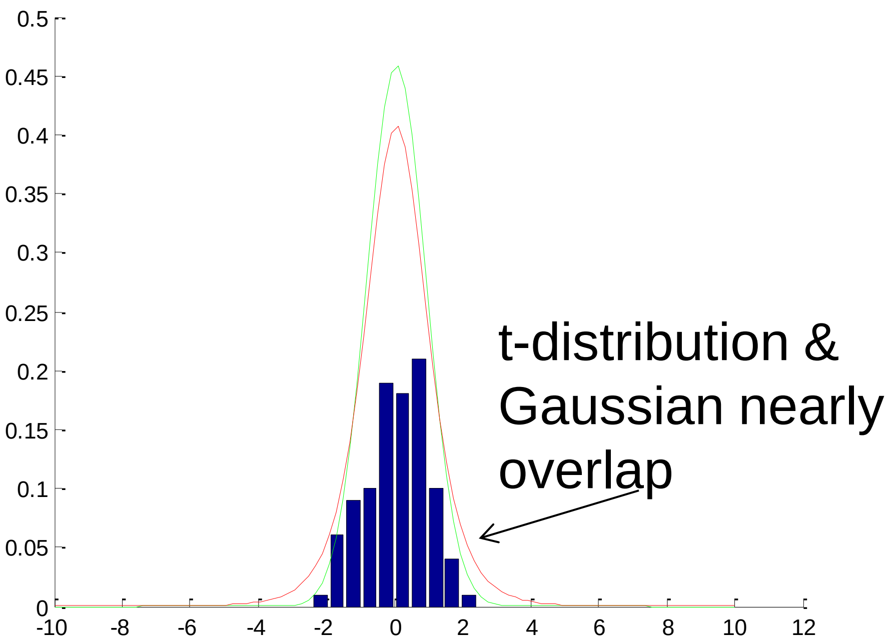
$$p(x \mid \mu, a, b) = \int_0^{\infty} \mathcal{N}\left(x \mid \mu, \tau^{-1}\right) \mathcal{Gamma}\left(\tau \mid a, b\right) d\tau =$$

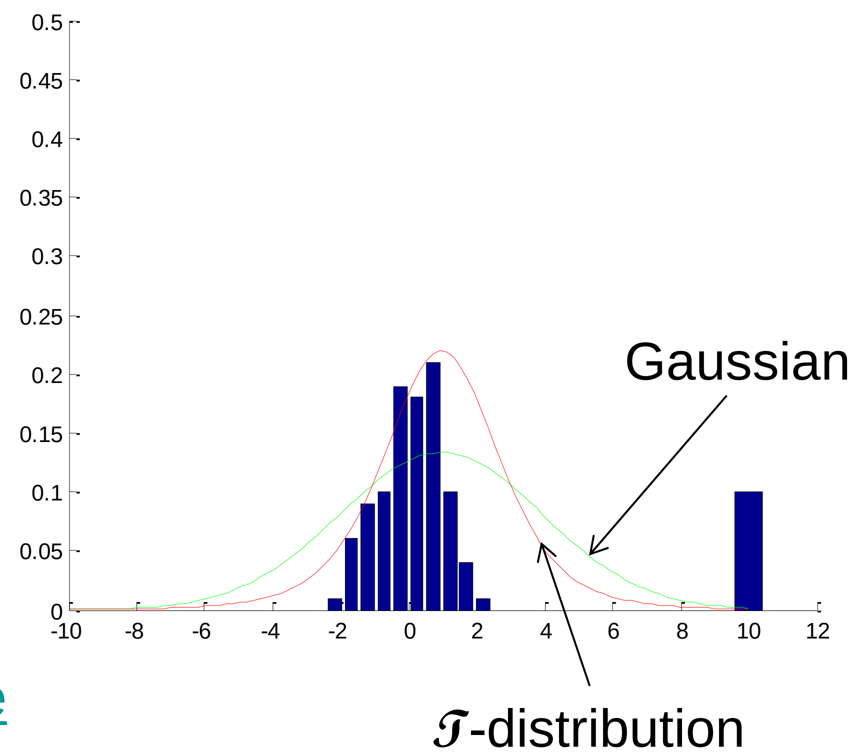
$$= \int_0^{\infty} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau}{2}\left(x-\mu\right)^2\right) \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau} d\tau$$

- The Student's $\mathcal{T}$ distribution can be seen from the equation above is a ***mixture of infinite Gaussian each of them with different precision***.
- The result is a distribution that in general has longer 'tails' than a Gaussian.
- This gives the $\mathcal{T}$-distribution robustness, i.e. the $\mathcal{T}$-distribution is much less sensitive than the Gaussian to the presence of outliers.

Robustness of Student's $\mathcal{T}$ Distribution

- The robustness of the $\mathcal{T}$-distribution is illustrated here by comparing the maximum likelihood solutions for a Gaussian and a $\mathcal{T}$-distribution (30 data points from the Gaussian are used). The effect of a small number of outliers is less significant for the $\mathcal{T}$-distribution than for the Gaussian.

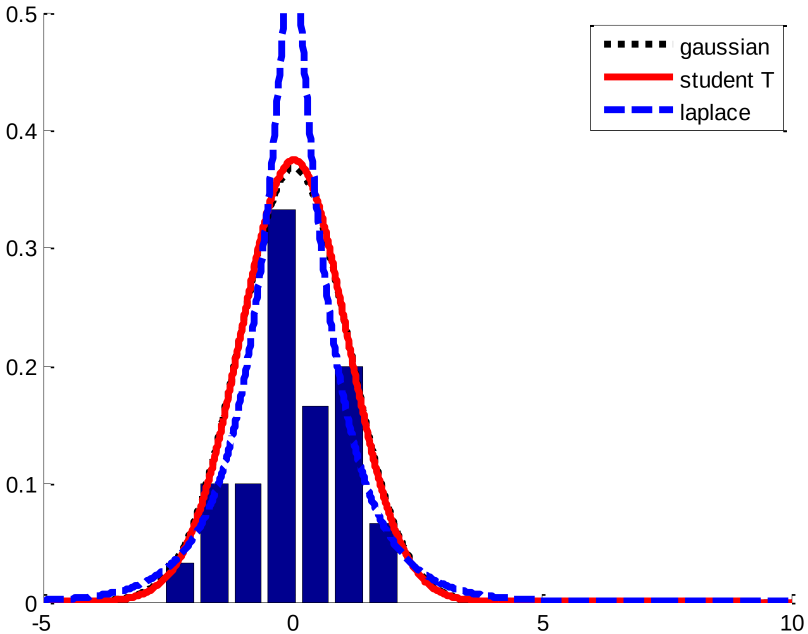


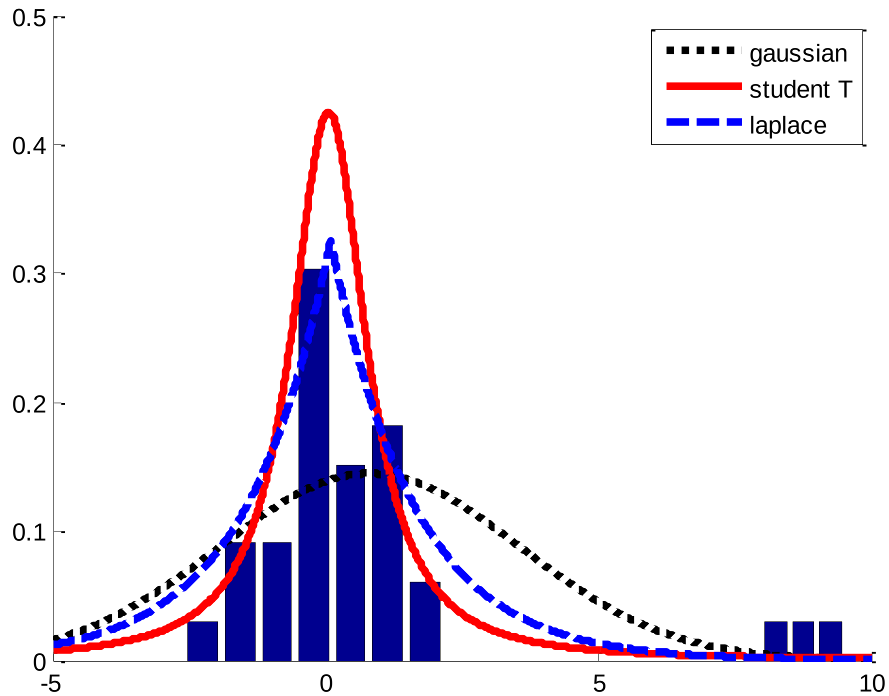


MatLab Code

Robustness of Student's $\mathcal{T}$ Distribution

- The earlier simulation is repeated here with the MPTK toolbox.





Run MatLab function *robustDemo*
from Kevin Murphys' PMTK

Multivariate Student's $\mathcal{T}$ Distribution

$$p(x \mid \mu, a, b) = \int_0^\infty \mathcal{N}\left(x \mid \mu, \tau^{-1}\right) \mathcal{Gamma}\left(\tau \mid a, b\right) d\tau$$

- If we return to the prior above and substitute $\upsilon = 2a,\ \lambda = \frac{a}{b},\ \eta = \tau b / a$, and use

$$\mathcal{Gamma}\left(\tau \mid a, b\right) = \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau}$$

we can write the Student's $\mathcal{T}$ distribution as :

$$\mathcal{T}(x \mid \mu, \lambda, \upsilon) = \int_0^\infty \mathcal{N}\left(x \mid \mu, (\eta\lambda)^{-1}\right) \mathcal{Gamma}\left(\eta \mid \upsilon/2, \upsilon/2\right) d\eta$$

- This form is useful in *providing generalization to a multivatiate Student's $\mathcal{T}$*

$$\mathcal{T}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \upsilon) = \int_0^\infty \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathcal{Gamma}\left(\eta \mid \upsilon/2, \upsilon/2\right) d\eta$$

Multivariate Student's $\mathcal{T}$ Distribution

$$\mathcal{T}(\mathbf{x}\,|\,\boldsymbol{\mu},\Lambda,\upsilon)=\int_0^\infty \mathcal{N}\left(\mathbf{x}\,|\,\boldsymbol{\mu},(\eta\Lambda)^{-1}\right)\mathcal{Gamma}\left(\eta\,|\,\upsilon/2,\upsilon/2\right)d\eta$$

- This integral can be computed analytically as:

$$\mathcal{T}(\mathbf{x}\,|\,\boldsymbol{\mu},\Lambda,\upsilon)=\frac{\Gamma(\frac{D}{2}+\frac{\upsilon}{2})}{\Gamma(\frac{\upsilon}{2})}\frac{|\Lambda|^{1/2}}{(\pi\upsilon)^{D/2}}\left[1+\frac{\Delta^2}{\upsilon}\right]^{-\upsilon/2-D/2}$$

$$\Delta^2=(\mathbf{x}-\boldsymbol{\mu})^T\Lambda(\mathbf{x}-\boldsymbol{\mu})\ (\textit{Mahalanobis Distance})$$

- One can derive the above form of the distribution by substitution in the Eq. on the top.

$$\mathcal{T}(\mathbf{x}\,|\,\boldsymbol{\mu},\Lambda,\upsilon)=\frac{(\upsilon/2)^{\upsilon/2}}{\Gamma(\upsilon/2)}\frac{|\Lambda|^{1/2}}{(2\pi)^{D/2}}\int_0^\infty \eta^{D/2}\eta^{\upsilon/2-1}e^{-\upsilon\eta/2}e^{-\eta\Delta^2/2}d\eta \qquad Use\ \tau=\eta\left(\upsilon/2+\Delta^2/2\right)$$

$$=\frac{(\upsilon/2)^{\upsilon/2}}{\Gamma(\upsilon/2)}\frac{|\Lambda|^{1/2}}{(2\pi)^{D/2}}\left(\upsilon/2+\Delta^2/2\right)^{-D/2-\upsilon/2}\int_0^\infty \tau^{D/2+\upsilon/2-1}e^{-\tau}d\tau=\frac{\Gamma(\upsilon/2+d/2)}{\Gamma(\upsilon/2)}\frac{|\Lambda|^{1/2}}{(\pi\upsilon)^{D/2}}\left(1+\Delta^2/\upsilon\right)^{-D/2-\upsilon/2}$$

- Normalization proof is immediate from the normalization of the normal & Gamma distributions

Multivariate Student's $\mathcal{T}$ Distribution

$$\mathcal{T}(\mathbf{x}\,|\,\boldsymbol{\mu},\Lambda,\upsilon)=\frac{\Gamma(\frac{D}{2}+\frac{\upsilon}{2})}{\Gamma(\frac{\upsilon}{2})}\frac{|\Lambda|^{1/2}}{(\pi\upsilon)^{D/2}}\left[1+\frac{\Delta^2}{\upsilon}\right]^{-\upsilon/2-D/2}$$

- Some useful results of the multivatiate Student's $\mathcal{T}$ are given below:

$$\mathbb{E}[\boldsymbol{x}]=\boldsymbol{\mu}\quad if\ \upsilon>1,\ cov[\boldsymbol{x}]=\frac{\upsilon}{\upsilon-2}\boldsymbol{\Lambda}^{-1}\quad if\ \upsilon>2,\ mode[\boldsymbol{x}]=\boldsymbol{\mu}$$

- One can show easily the expression for the mean by using **x**=**z**+$\boldsymbol{\mu}$:

$$\mathbb{E}[\boldsymbol{x}]=\int_{-\infty}^{+\infty}\frac{\Gamma(\frac{D}{2}+\frac{\upsilon}{2})}{\Gamma(\frac{\upsilon}{2})}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\upsilon)^{D/2}}\left[1+\frac{\Delta^2}{\upsilon}\right]^{-\upsilon/2-D/2}(\boldsymbol{z}+\boldsymbol{\mu})\,d\boldsymbol{z}$$

- The 1st term drops out since $\mathcal{T}(\boldsymbol{z}\,|\,\boldsymbol{0},\Lambda,\upsilon)$ is even. The 2nd term gives $\boldsymbol{\mu}$ from the normalization of the distribution.

- The covariance is computed as:

$$\mathrm{cov}[\boldsymbol{x}]=\int_{\eta=0}^{+\infty}\left[\int_{\boldsymbol{x}}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\Lambda)^{-1}\right)(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\,d\boldsymbol{x}\right]\mathcal{Gamma}(\eta\,|\,\upsilon/2,\upsilon/2)\,d\eta=\frac{(\upsilon/2)^{\upsilon/2}}{\Gamma(\upsilon/2)}\int_{\eta=0}^{+\infty}(\eta\boldsymbol{\Lambda})^{-1}\eta^{\upsilon/2-1}e^{-\upsilon/2\eta}d\eta$$

$$=\boldsymbol{\Lambda}^{-1}\frac{(\upsilon/2)^{\upsilon/2}}{\Gamma(\upsilon/2)}\frac{\Gamma(\upsilon/2-1)}{(\upsilon/2)^{\upsilon/2}(\upsilon/2)^{\upsilon/2-1}}=\frac{(\upsilon/2)\Gamma(\upsilon/2-1)}{\Gamma(\upsilon/2)}\boldsymbol{\Lambda}^{-1}=\frac{\upsilon/2}{\upsilon/2-1}\boldsymbol{\Lambda}^{-1}=\frac{\upsilon}{\upsilon-2}\boldsymbol{\Lambda}^{-1}$$

Multivariate Student's $\mathcal{T}$ Distribution

$$\mathcal{T}(\mathbf{x}\,|\,\boldsymbol{\mu},\Lambda,\upsilon)=\frac{\Gamma\left(\frac{D}{2}+\frac{\upsilon}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{\left(\pi\upsilon\right)^{D/2}}\left[1+\frac{\Delta^2}{\upsilon}\right]^{-\upsilon/2-D/2}$$

- Differentiation with respect to x also shows the mode being $\boldsymbol{\mu}$:

$$\mathbb{E}[\boldsymbol{x}]=\boldsymbol{\mu}\quad if\ \ \upsilon>1,\ cov[\boldsymbol{x}]=\frac{\upsilon}{\upsilon-2}\boldsymbol{\Lambda}^{-1}\quad if\ \ \upsilon>2,\ mode[\boldsymbol{x}]=\boldsymbol{\mu}$$

- The Student's $\mathcal{T}$ has fatter tails than a Gaussian. *The smaller ν is the fatter the tails.*

- For $\nu\to\infty$, the distribution approaches a Gaussian. Indeed note that:

$$\left[1+\frac{\Delta^2}{\upsilon}\right]^{-\upsilon/2-D/2}=\exp\left(-\left(\frac{\upsilon}{2}+\frac{D}{2}\right)\ln\left[1+\frac{\Delta^2}{\upsilon}\right]\right)\overset{\upsilon\to\infty}{=}\exp\left(-\frac{\upsilon}{2}\left(\frac{\Delta^2}{\upsilon}-\frac{1}{2}\left(\frac{\Delta^2}{\upsilon}\right)^2\right)\right)=\exp\left(-\frac{\Delta^2}{2}+O\left(\upsilon^{-1}\right)\right)$$

- The distribution can also be written in terms of $\boldsymbol{\Sigma}=\boldsymbol{\Lambda}^{-1}$ (scale matrix – not the covariance) or $\mathbf{V}=\nu\boldsymbol{\Sigma}$.

The Laplace Distribution

- Another distribution with heavy tails is the Laplace distribution, also known as the *double sided exponential distribution*. It has the following pdf:

$$\mathcal{Lap}\left(x \mid \mu, b\right) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

- μ is a location parameter and b > 0 is a scale parameter

$$Mean = \mu,\ Mode = \mu, Var = 2b^2$$

- Its robust to outliers (see an earlier demonstration).

- *It puts mores probability density at 0 than the Gaussian. This property is a useful way to encourage sparsity in a model*.