A Mixture of Gaussian and Student's t Errors for a Robust and Accurate Inference

Hyungsuk Tak

SAMSI / UNC Chapel Hill

28 Feb 2017

Motivation

- Gaussian error, $\epsilon_j \sim \mathrm{N}_1(0,\ V_j)$

  - Pros : (Thin tails) Efficient inference when $\nexists$ outliers.
  - Cons: Can bias inference when $\exists$ outliers.

- Student's t_ν error, $\epsilon_j \sim V_j^{0.5} t_\nu$

  - Pros: (Heavy tails) Robust to outliers.
  - Cons: Can be less efficient due to unnecessarily heavy tails for most of the normally observed data.

- Why not use both? A mixture of Gaussian and Student's t_ν errors,

$$\begin{aligned} \epsilon_j &\sim \mathrm{N}_1(0,\ V_j) \quad \text{with probability } 1-\theta_j, \\ &\sim V_j^{0.5} t_\nu \qquad\quad \text{with probability } \theta_j, \end{aligned}$$

enabling a robust & accurate estimation.

The mixture error

A p-dimensional mixture error with known variance component $\mathbf{V}_j$ is

$$\begin{aligned} \boldsymbol{\epsilon}_j \mid z_j, \alpha_j &\sim \mathrm{N}_p(\mathbf{0},\ {\alpha_j}^{z_j}\mathbf{V}_j), \\ z_j \mid \theta_j &\sim \mathrm{Bernoulli}(\theta_j), \\ \theta_j &\sim \mathrm{Uniform}(0,1), \\ \alpha_j &\sim \mathrm{Inv.Gamma}(\nu/2,\ \nu/2). \end{aligned}$$

- z_j, a latent outlier indicator.
- θ_j, probability of datum j being an outlier.
- α_j, an auxiliary variable to express t_ν as a scale mixture of Gaussian.

  (If $x \mid \alpha \sim \mathrm{N}(0, \alpha)$ & $\alpha \sim \mathrm{Inv.Gamma}(\nu/2, \nu/2)$, then $x \sim t_\nu$.)

  (West, 1987; Peel and McLachlan, 2000; Gelman et al., 2013)

Relationship with other errors

$$\begin{aligned}
\boldsymbol{\epsilon}_j \mid z_j, \alpha_j &\sim \mathrm{N}_p(\mathbf{0},\ \alpha_j^{z_j}\mathbf{V}_j),\\
z_j \mid \theta_j &\sim \mathrm{Bernoulli}(\theta_j),\\
\theta_j &\sim \mathrm{Uniform}(0,1),\\
\alpha_j &\sim \text{inverse-Gamma}(\nu/2,\ \nu/2).
\end{aligned}$$

This proposed mixture error is

- marginally (α_j & θ_j) equivalent to a Gaussian error if $z_j = 0$ for all j.
- marginally (α_j & θ_j) equivalent to a t_ν error if $z_j = 1$ for all j.
- marginally (α_j & z_j) equivalent to a mixture of Gaussian & t_ν errors.
- marginally (z_j) equivalent to a mixture of two Gaussians if α_j is fixed at a constant (e.g., MLE) (Aitkin and Wilson, 1980; Hogg et al., 2010; Vallisneri and van Haasteren, 2016).

Converting Gaussian error to mixture error

Simply multiplying $\alpha_j^{z_j}$ to V_j with prior distributions on the additional parameters converts Gaussian errors to mixture errors:

$$\text{From} \qquad \epsilon_j \sim \mathrm{N}_p(0,\ V_j)$$

$$\begin{aligned} \text{to} \quad \epsilon_j \mid z_j, \alpha_j &\sim \mathrm{N}_p(0,\ \alpha_j^{z_j} V_j), \\ z_j \mid \theta_j &\sim \mathrm{Bernoulli}(\theta_j), \\ \theta_j &\sim \mathrm{Uniform}(0, 1), \\ \alpha_j &\sim \mathrm{Inv.Gamma}(\nu/2,\ \nu/2). \end{aligned}$$

If $\exists$ a Gibbs sampler derived from a Gaussian error model, we can still use it after replacing V_j with $\alpha_j^{z_j} V_j$ to fit a mixture error model, additionally updating z_j (Bernoulli), θ_j (Beta), and α_j (Inv.Gamma).

Example 1: Unknown location

Three data sets:

- Original Data: $y_j \overset{i.i.d.}{\sim}$ N(0, 1) for $j = 1, 2, \ldots, 20$.
- Data with an outlier: The same data except $y_{20} = -10$ or $y_{20} = 10$.

Suppose the mean (μ) of the generative Gaussian distribution is unknown.

Three error models with an improper flat prior on μ:

- Gaussian error: $y_j \mid \mu = \mu + \epsilon_j, \;\; \epsilon_j \sim \text{N}(0,\ 1)$
- t_4 error (Chp 17, Gelman et al., 2013):

$$y_j \mid \mu = \mu + \epsilon_j, \;\; \epsilon_j \sim t_4$$

- Mixture error:

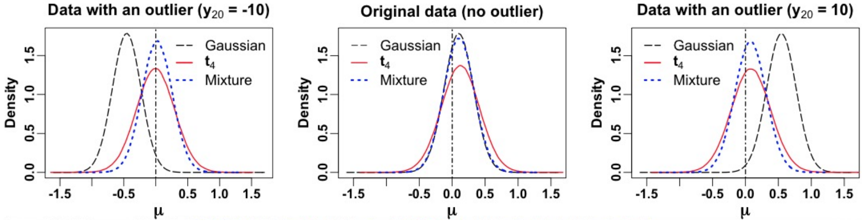
$$y_j \mid \mu, z_j, \alpha_j = \mu + \epsilon_j, \;\;\; \epsilon_j \sim N(0, \alpha_j^{z_j}),$$

$$z_j \sim \text{Bernoulli}(0.1),$$

$$\alpha_j \sim \text{Inv.Gamma}(2, 2).$$

Example 1: Unknown location (cont.)

Marginal posterior distribution of μ based on a million posterior samples. The generative value $\mu = 0$ is denoted by a vertical dot-dashed line.



- Without the outlier in the 2nd panel, the dotted curve (mixture) passes in-between the dashed (Gaussian) and solid (t_4) curves, a mixture effect.
- In the 1st and 3rd panels, the mixture error model maintains this mixture effect robustly when $\exists$ an outlier, enabling a robust and more accurate inference than t_4 error model.

Example 2: A hierarchical model

31 NYC hospital profiling data for CABG surgery (Morris and Lysy, 2012)

- Data: Index of successful surgery rate y_j with known variance V_j.

j	y_j	V_j
1	-2.07	2.78^2
2	-0.22	2.76^2
$\vdots$	$\vdots$	$\vdots$
31	1.14	0.62^2

- A two-level Gaussian model

  $y_j \mid \mu_j = \mu_j + \epsilon_j, \;\; \epsilon_j \sim \mathrm{N}_1(0,\ V_j),$

  $\mu_j \mid \beta, A \sim \mathrm{N}_1(\beta, A),$

  $\pi(\beta, A) \propto \exp\left(-\frac{\beta^2}{2\times 10^5}\right) \times \frac{I_{\{A>0\}}}{(10^5+A)^2},$

  where μ_j is random effect j, β and A are mean and variance of the prior distribution of μ_j.

Example 2: A hierarchical model (cont.)

- Simulated data (with known parameter values)
  - Simulated data 1: Generate μ_j's given $\beta = 0$ and $A = 0.722$, and then generate y_j^*'s given the simulated μ_j's.
  - Simulated data 2: The same as y_j^*'s except one synthetic outlier, setting $y_1^* \leftarrow y_1^* + 4V_1^{0.5}$
  - Simulated data 3: The same as y_j^*'s except three synthetic outliers, with y_1^*, y_2^*, and y_3^* similarly manipulated.

- A mixture error model with additional parameters in red color

$$y_j \mid \mu_j, z_j, \alpha_j = \mu_j + \epsilon_j, \;\; \epsilon_j \sim \mathrm{N}_1(0, \alpha_j^{z_j} V_j),$$

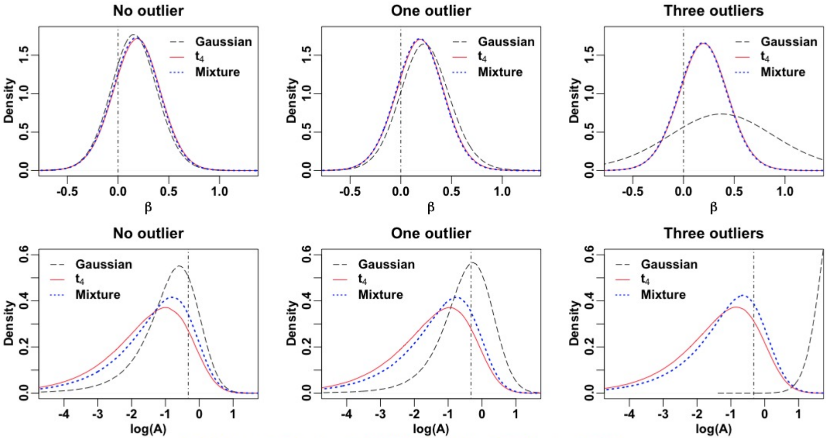
$$\mu_j \mid \beta, A \sim \mathrm{N}_1(\beta, A),$$

$$\pi(\beta, A) \propto \exp\left(-\frac{\beta^2}{2V_0}\right) \times \frac{I_{\{A>0\}}}{(V_0+A)^2},$$

$$z_j \mid \theta_j \sim \text{Bernoulli}(\theta_j),\; \theta_j \sim \text{Uniform}(0,1),\; \alpha_j \sim \text{Inv.Gamma}(2,2).$$

Example 2: A hierarchical model (cont.)

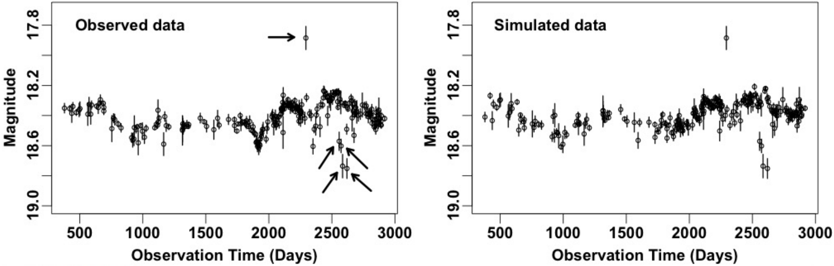
Marginal posterior distribution of β is in the first row and that of $\log(A)$ is in the second row for each case (column) from 15 million posterior draws.



The mixture effect, clearly shown in the first column, is maintained robustly when $\exists$ outlier(s) in the second and third columns, enabling a robust and accurate inference.

Example 3: A damped random walk process

A MACHO (Massive Compact Halo Objects) quasar light curve[1] (Geha et al., 2003), observed via a V-band optical filter on 246 nights for 7.5 years since 1992, and its simulated light curve (right panel).



[1] http://www.astro.yale.edu/mgeha/MACHO/63.7365.151.html

Example 3: A damped random walk process

A Gaussian error model based on a damped random walk process (Kelly et al., 2009) has been used to fit a quasar light curve: ($j = 1, 2, \ldots, 246$)

$$y_j \mid Y(t_j) = Y(t_j) + \epsilon_j, \quad \epsilon_j \sim \mathrm{N}_1(0,\ V_j),$$

$$Y(t_1) \mid \mu, \sigma^2, \tau \ \sim\ \mathrm{N}_1\left(\mu,\ \frac{\tau\sigma^2}{2}\right), \text{ and for } i = 2, 3, \ldots, 246,$$

$$Y(t_i) \mid Y(t_{i-1}), \mu, \sigma^2, \tau \ \sim\ \mathrm{N}_1\left(\mu + a_j\big(Y(t_{i-1}) - \mu\big),\ \frac{\tau\sigma^2}{2}(1 - a_i^2)\right),$$

$$\mu \sim \mathrm{Uniform}(-30,\ 30),$$

$$\sigma^2 \sim \text{inverse-Gamma}(1,\ 2 \times 10^{-7}),$$

$$\tau \sim \text{inverse-Gamma}(1,\ 1).$$

Sampling the full posterior, $\pi(\mathbf{Y}(\mathbf{t}), \mu, \sigma^2, \tau \mid \mathbf{y})$, via Gibbs sampling.

Example 3: A damped random walk process

Deriving a mixture error model from a Gaussian error model:

1. Multiply $\alpha_j^{z_j}$ to V_j, i.e., for $j = 1, 2, \ldots, 246$,

$$y_j \mid Y(t_j), \alpha_j, z_j = Y(t_j) + \epsilon_j, \quad \epsilon_j \sim \mathrm{N}_1(0,\ \alpha_j^{z_j} V_j),$$

$$Y(t_1) \mid \mu, \sigma^2, \tau \ \sim\ \mathrm{N}_1\left(\mu,\ \frac{\tau\sigma^2}{2}\right), \text{ and for } i = 2, 3, \ldots, 246,$$

$$Y(t_i) \mid Y(t_{i-1}), \mu, \sigma^2, \tau \ \sim\ \mathrm{N}_1\left(\mu + a_j\big(Y(t_{i-1}) - \mu\big),\ \frac{\tau\sigma^2}{2}(1 - a_i^2)\right),$$

$$\mu \sim \mathrm{Uniform}(-30,\ 30),$$

$$\sigma^2 \sim \text{inverse-Gamma}(1,\ 2 \times 10^{-7}),$$

$$\tau \sim \text{inverse-Gamma}(1,\ 1),$$

2. Set prior distributions for the additional parameters

$$z_j \mid \theta_j \sim \mathrm{Bernoulli}(\theta_j),\ \theta_j \sim \mathrm{Uniform}(0, 1),\ \alpha_j \sim \mathrm{Inv.Gamma}(2, 2).$$

Example 3: A damped random walk process

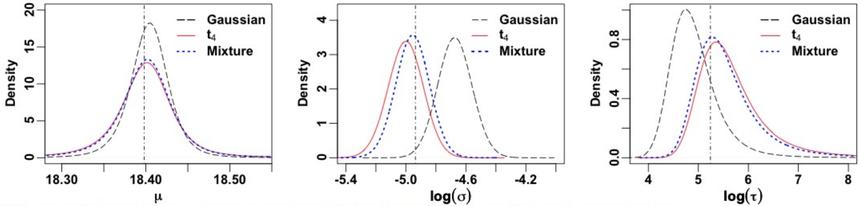
Sampling the extended full posterior, $\pi^*(\mathbf{Y}(\mathbf{t}), \mu, \sigma^2, \tau, \boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y})$:

1. Sampling $\mathbf{Y}(\mathbf{t}), \mu, \sigma^2, \tau$ via the original Gibbs sampler, originally used to sample $\pi(\mathbf{Y}(\mathbf{t}), \mu, \sigma^2, \tau \mid \mathbf{y})$, after replacing V_j with $\alpha_j^{z_j} V_j$.
2. Sampling additional parameters,
   z_j (Bernoulli), θ_j (Beta), and α_j (Inv.Gamma) for all j.

We independently run 30 Markov chains each for 500,000 iterations and then combine these chains to summarize the sampling results.

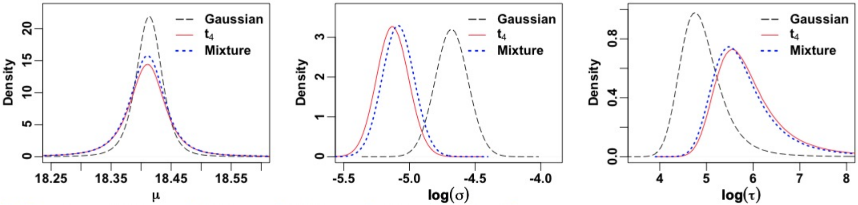
Example 3: A damped random walk process

Simulated data (vertical lines indicates generative values)



Posterior distributions obtained with mixture errors (dotted) concentrate more on the generative values than those obtained with t_4 errors (solid).

MACHO source 63.7365.151 data



Example 3: A damped random walk process

Table 1 : (Simulation) Mode estimate is the average of the thirty posterior modes, bias, mean squared error (MSE) ratio defined as the MSE obtained with the Gaussian or t_4 error divided by that obtained with the mixture error, 95% quantile-based posterior interval (P.I.) and its length computed from the combined fifteen million posterior samples, and average CPU time in seconds.

	Error	Mode	Bias	MSE Ratio	95% P.I.	Length of P.I.	CPU time
	N	18.406	0.00825	0.41	(18.342, 18.462)	0.119	2695
μ	t_4	18.396	0.00180	0.78	(18.285, 18.494)	0.209	2744
	N+t_4	18.405	0.00747	-	(18.294, 18.490)	0.195	2982
	N	-4.674	0.26244	179.30	(-4.895, -4.447)	0.448	2695
$\log(\sigma)$	t_4	-4.999	0.06250	10.20	(-5.235, -4.772)	0.463	2744
	N+t_4	-4.955	0.01893	-	(-5.177, -4.737)	0.440	2982
	N	4.746	0.49522	91.08	(4.191, 6.151)	1.960	2695
$\log(\tau)$	t_4	5.351	0.10970	4.77	(4.739, 7.617)	2.878	2744
	N+t_4	5.284	0.04182	-	(4.681, 7.421)	2.740	2982

Conclusion

A mixture error model results in a robust and more accurate parameter estimation in the presence of outliers than a t_4 error model (anecdotal evidence).

It is simple and always possible to convert a Gaussian error to a mixture error by multiplying $\alpha_j^{z_j}$ to the (known) variance component V_j.

Computational cost for additional parameters is not expensive.

References

1. Aitkin, M. and Wilson, G. T. (1980) "Mixture Models, Outliers, and the EM Algorithm" *Technometrics*, **22**, 3, 325–331.
2. Geha et al. (2003) "Variability-Selected Quasars in MACHO Project Magellanic Cloud Fields" *The Astrophysical Journal*, **125**, 1, 1.
3. Gelman et al. (2013). Bayesian Data Analysis. *CRC press*.
4. Kelly et al. (2009) "Are the Variations in Quasar Optical Flux Driven by Thermal Fluctuations?" *The Astrophysical Journal*, **698**, 1, 895.
5. Morris, C. N. and Lysy, M. (2012) "Shrinkage Estimation in Multilevel Normal Models" *Statistical Science*, **27**, 1, 115–134.
6. Peel et al. (2000) "Robust Mixture Modelling using the t Distribution" *Statistics and Computing*, **10**, 4, 339–348.
7. Hogg et al. (2010) "Data Analysis Recipes: Fitting a Model to Data" *arXiv:1008.4686*.
8. Vallisneri, M. and van Haasteren, R. (2016) "Taming Outliers in Pulsar-Timing Datasets with Hierarchical Likelihoods and Hamiltonian Sampling" *arXiv:1609.02144*.
9. West, M. (1987) "On Scale Mixtures of Normal Distributions" *Biometrika*, **74**, 3, 646–648.