THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

Putting it all together

Sinead Williamson
The University of Texas Department of Statistics and Data Science

Starting point: Bayesian linear regression

Basic model:

$$\begin{aligned}
\mathbf{y}|\beta, X \sim & \mathsf{Normal}(X\beta, (\omega\Lambda)^{-1}) \\
\beta \sim & \mathsf{Normal}(\mu, (\omega K)^{-1}) \\
\omega \sim & \mathsf{Gamma}(a, b)
\end{aligned}$$

Let's look at what this looks like... [notebook]

We can modify this in a wide variety of ways! Some of which we've played around with...

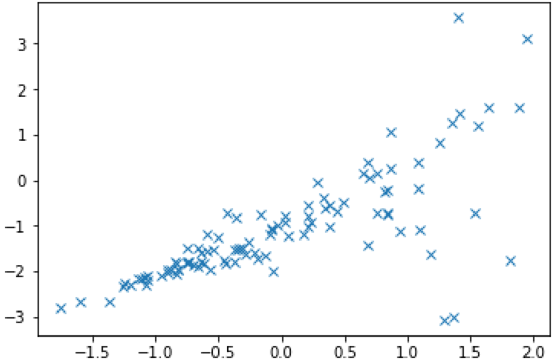
- Switch out likelihood for a different distribution
- Allow variance to vary between individuals $\rightarrow$ heavy tails
- Allow variance to vary between groups

We can modify this in a wide variety of ways! Some of which we've played around with...

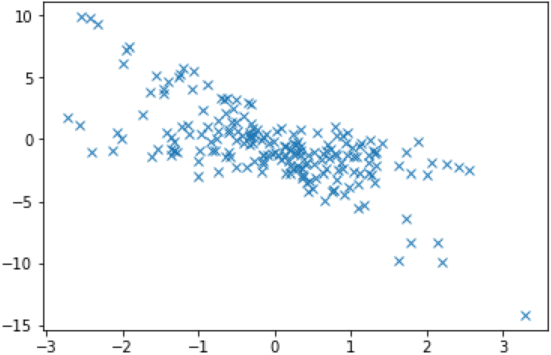
- Switch out likelihood for a different distribution
- Allow variance to vary between individuals $\rightarrow$ heavy tails
- Allow variance to vary between groups

Some of which we haven't!

How could we model data that looks like this?



How about data that looks like this?

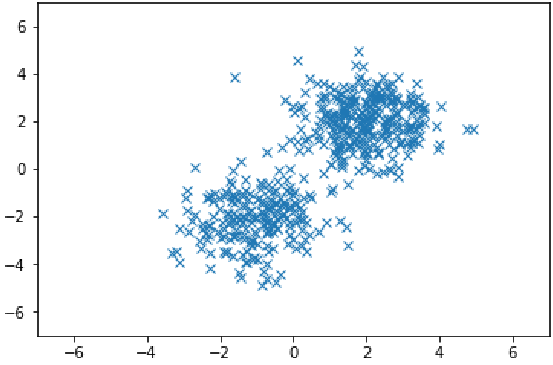


What could we do if we were missing covariates?

X1	X2	X3	Y
1.45	0.22	0.73	3.88
0.62	-	1.21	1.56
2.21	1.67	1.08	3.42

The extreme version of missing (categorical) covariates is a Gaussian mixture model:

$$\begin{aligned}
\pi &\sim \text{Dirichlet}(\alpha) \\
Z_i &\sim \pi \\
\mu_k &\sim \mathit{Normal}(\mu_0, \sigma_0^2) \\
\omega_k &\sim \mathit{Gamma}(a, b) \\
X_i &\sim \mathit{Normal}(\mu_{Z_k}, 1/\omega_{Z_k})
\end{aligned}$$



$$\pi \sim \text{Dirichlet}(\alpha) \qquad Z_i \sim \pi$$

$$\mu_k \sim Normal(\mu_0, \sigma_0^2) \qquad \omega_k \sim Gamma(a, b) \qquad X_i \sim Normal(\mu_{Z_k}, 1/\omega_{Z_k})$$

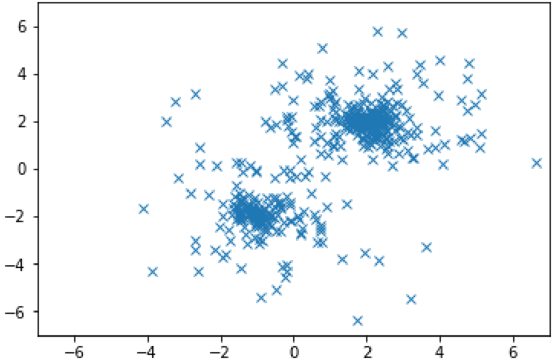
- Conditioned on Z_i, we have a linear regression model. We can either sample ω and μ, or integrate them out.
- Conditioned on π, ω and μ, we have

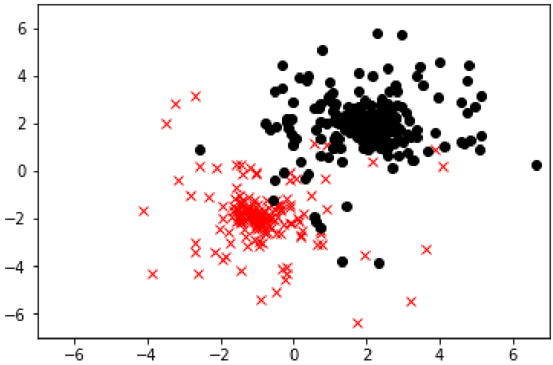
$$P(Z_i = k|\theta, \mu, \omega) \propto \pi_k N(X_i; \mu_k, 1/\omega_k)$$

  - Can construct a normalized vector $\hat{p} : \hat{p}_k = \frac{\pi_k N(X_i;\mu_k,1/\omega_k)}{\sum_j \pi_j N(X_i;\mu_j,1/\omega_j)}$
  - Can then sample from a multinomial with probability $\hat{p}$.
- Can integrate out π:

$$P(Z_i = k|\theta, \mu, \omega) \propto (\sum_{j \neq i} I(Z_j = k))N(X_i; \mu_k, 1/\omega_k)$$

What could we do if our data looked like this?





6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

If we get a mixture model by putting a prior over latent categorical regressors...

What do we get if we put a prior over latent continuous valued regressors?

Regression:

$$y|X, \beta \sim \text{Normal}(X\beta^T, \sigma^2)$$

Multivariate extension:

$$Y|X, \beta \sim \text{Normal}(X\beta^T, \sigma^2 I)$$

Regression:

$$y|X,\beta \sim \text{Normal}(X\beta^T, \sigma^2)$$

Multivariate extension:

$$Y|X,\beta \sim \text{Normal}(X\beta^T, \sigma^2 I)$$

Replace observed X with latent variables F...

$$\begin{aligned} F &\sim \text{Normal}(0, \sigma_Z^2 I) \\ \Lambda &\sim \text{Normal}(0, \sigma_\Lambda^2) \\ Y|F,\Lambda &\sim \text{Normal}(F\Lambda^T, \sigma_Y^2) \end{aligned}$$

What is $p(y_i|\Lambda, \sigma_Y^2)$?

$$
\begin{aligned}
p(y_i|\Lambda) &= \int p(y_i|f_i,\Lambda)p(f_i)df_i \\
&\propto \int \exp\left\{-\frac{1}{2\sigma_y^2}(y_i-\Lambda f_i)^T(y_i-\Lambda f_i)\right\}\exp\left\{-\frac{1}{2}f_i^T f_i\right\}df_i \\
&= \int \exp\left\{-\frac{1}{2}\left((y_i-\Lambda f_i)^T(y_i-\Lambda f_i)/\sigma_y^2+f_i^T f_i\right)\right\} \\
&= \int \exp\left\{-\frac{1}{2}\left(\frac{y_i^T y_i - f_i^T\Lambda^T y_i - y_i^T\Lambda f_i + f_i^T\Lambda\Lambda^T f_i}{\sigma_y^2} f_i^T f_i\right)\right\}df_i \\
&= \int \exp\left\{-\frac{1}{2}(f_i-m)^T\Sigma^{-1}(f_i-m)+y_i^T(\sigma_y^2 I+\Lambda\Lambda^T)^{-1}y_i\right\}df_i \\
&\quad \left(\text{where } \Sigma=(I+\Lambda\Lambda^T/\sigma_y^2)^{-1}, m=\Sigma\Lambda^T y_i/\sigma_y^2\right) \\
&\propto \exp\left\{-\frac{1}{2}y_i^T(\sigma_y^2 I+\Lambda\Lambda^T)^{-1}y_i\right\}
\end{aligned}
$$

So, $y_i|\Lambda,\sigma^2 \sim \text{Normal}(0,\sigma_y I+\Lambda\Lambda^T)$

For a Gibbs sampler, we need the conditional distributions $p(F|\Lambda, Y)$ and $p(\Lambda|F, Y)$.

$$\begin{aligned} p(\lambda_d|y, F) \propto & p(y^d|F, \Lambda_d)p(\Lambda_d) \\ & \propto \exp\left\{-\frac{1}{2}\left(\frac{(y^d - F\Lambda_d)^T(y^d - F\Lambda_d)}{\sigma_y^2} - \frac{\Lambda_k^T\Lambda_k}{\sigma_\Lambda^2}\right)\right\} \end{aligned}$$

For a Gibbs sampler, we need the conditional distributions $p(F|\Lambda, Y)$ and $p(\Lambda|F, Y)$.

$$\begin{aligned} p(\lambda_d|y, F) &\propto p(y^d|F, \Lambda_d)p(\Lambda_d) \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{(y^d - F\Lambda_d)^T(y^d - F\Lambda_d)}{\sigma_y^2} - \frac{\Lambda_k^T\Lambda_k}{\sigma_\Lambda^2}\right)\right\} \end{aligned}$$

Look familiar?? Conditioned on F, we have a linear regression model!

$$\lambda_d|y^d, F \sim \text{Normal}(m, S)$$

where

$$S = \left(\sigma_\lambda^{-2} I + \sigma_y^{-2} F^T F\right)^{-1} m = \quad SF^T y^d$$

What about $p(f_i|\Lambda, Y)$?

What about $p(f_i|\Lambda, Y)$?

Well, our model is symmetric... conditioned on Λ, we have what looks like a regression:

$$p(f_i|y_i, \Lambda) \propto p(y_i|f_i, \Lambda)p(f_i)$$
$$\propto \exp\left\{-\frac{1}{2}\left(\frac{(y_i - \Lambda f_i)^T(y_i - \Lambda f_i)}{\sigma_y^2} + f_i^T f_i\right)\right\}$$

So,

$$f_i \sim \text{Normal}(m, S)$$

where

$$S = \left(I + \sigma_y^{-2}\Lambda^T\Lambda\right)^{-1} m = \quad S\Lambda^T y_i$$

What about $p(f_i|\Lambda, Y)$?

Well, our model is symmetric... conditioned on Λ, we have what looks like a regression:

$$p(f_i|y_i, \Lambda) \propto p(y_i|f_i, \Lambda)p(f_i)$$
$$\propto \exp\left\{-\frac{1}{2}\left(\frac{(y_i - \Lambda f_i)^T(y_i - \Lambda f_i)}{\sigma_y^2} + {f_i}^T f_i\right)\right\}$$

So,

$$f_i \sim \text{Normal}(m, S)$$

where

$$S = \left(I + \sigma_y^{-2}\Lambda^T\Lambda\right)^{-1} m = \quad S\Lambda^T y_i$$

Let's take a look at what this looks like! [notebook]

- We saw, in our outputs, evidence of a lack of identifiability.
- The two solutions are equally "good".
- If we have an orthogonal transform H such that $HH^T = H^T H = I$, then we can write

$$Y = \Lambda F + \epsilon \Lambda H H^T F + \epsilon = \tilde{\Lambda}\tilde{F} + \epsilon$$

- F and $\tilde{F}$ have the same statistical properties:

$$\begin{aligned} E[\tilde{F}] =& H^T E[F] = 0 \\ cov(\tilde{F}) =& H^T cov(F) H = H^T H = I \end{aligned}$$

Netflix problem...

	Iron Man	Avengers	The Notebook	It	Saw
James	3	2	1	5	5
Joe	4	5	4	1	1
Anna	1	?	2	5	4
Beth	4	4	3	2	1

How could we model this?