

Bayesian Factorization Machines with Stan and R

Adam M. Lauretig

12/18/2019

Welcome

- ▶ Talk: Factorization machines:
 - ▶ What are they?
 - ▶ How do we fit them?
- ▶ Stan code + simulations!
 - ▶ Start with simple model
 - ▶ Next, hierarchical model

Who I Am

- ▶ Senior Data Scientist, JUST Capital
- ▶ Previously, Ph.D. in political science (at The Ohio State University)
- ▶ Word Embeddings, Causal Inference, Bayesian Statistics, Discrete Choice Models

Why We Model

- ▶ Descriptive Inference: is there an association between \mathbf{X} and \mathbf{y} ?
- ▶ Prediction: with \mathbf{X} and \mathbf{y} , what is \mathbf{y} for a new \mathbf{X} ?
- ▶ Causal Inference: does changing \mathbf{X} , change \mathbf{y} ?
- ▶ Shameless plug: Book Chapter **[insert link](#)**

Latent Variable Models/Unsupervised models

- ▶ Learning parameters to reconstruct observed data
- ▶ Ex: Principal Components Analysis, Factor Analysis, Matrix Factorization, etc
- ▶ Data $\mathbf{X}_{N \times J}$ is decomposed into two low rank matrices: $\boldsymbol{\gamma}_{N \times K}$ and $\boldsymbol{\delta}_{K \times J}$
- ▶ Various assumptions about the structure of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$

Factorization Machines

- ▶ Combine regression with a latent variable model on the residuals

But Why?

Factorization Machines

- ▶ Regression Model, for one observation
- ▶ Categorical predictors $\mathbf{x}_{n \in N}$, $\mathbf{x}_{j \in J}$
- ▶ Outcome \mathbf{y}
- ▶ Parameters β

$$y_{nj} = \mathbf{x}_n \beta_1 + \mathbf{x}_j \beta_2 + \varepsilon_{nj}$$

```
m1 <- lm(y ~ factor(group_1) + factor(group_2))
```

- ▶ With Interactions

$$y_{nj} = \mathbf{x}_n \beta_1 + \mathbf{x}_j \beta_2 + \mathbf{x}_n \times \mathbf{x}_j \beta_3 + \varepsilon_{nj}$$

```
m1 <- lm(y ~ factor(group_1) + factor(group_2) + factor(group_1) * factor(group_2))
```


Factorization Machines

Problems:

- ▶ We can only estimate β_3 for observed interactions
- ▶ As N and J grow, β_3 increases with size $N * J$

Factorization Machines

Solution!

- ▶ Replace β_3 with the dot product of low-rank latent factors:
- ▶ $\gamma_{N \times K}$
- ▶ $\delta_{J \times K}$
- ▶ β_3 is now $\gamma_n \cdot \delta_j^\top$

Factorization Machines

- ▶ Interaction model:

$$y_{nj} = \mathbf{x}_n \beta_1 + \mathbf{x}_j \beta_2 + \boldsymbol{\gamma}_n \cdot \boldsymbol{\delta}_j^\top + \varepsilon_{nj}$$

- ▶ Depending on our assumptions about $\boldsymbol{\gamma}_n \cdot \boldsymbol{\delta}_j^\top$, we can now create FMs

Factorization Machines

Basic model - Each element of δ_j and γ_n is distributed standard normal: $\mathcal{N}(0, 1)$ - Automatic Relevance Determination (ARD) prior to shrink the matrix rank

$$y_{nj} \sim \mathcal{N}(\mathbf{x}_n \beta_1 + \mathbf{x}_j \beta_2 + \gamma_n \cdot \delta_j^\top, 1)$$

$$\beta \sim \mathcal{N}(0, \sigma^2)$$

$$\gamma_{n,k} \sim \mathcal{N}(0, 1)$$

$$\delta_{j,k} \sim \mathcal{N}(0, 1)$$

Simulating Data:

First, the regression component:

```
seed_to_use = 123
N = 1
J = 1
K = 1
set.seed(seed_to_use)
# number of levels for first covariate
N <- N
group_1 <- paste0("i", 1:N)
# number of levels for second covariate
J <- J
group_2 <- paste0("j", 1:J)
# number of latent dimensions
K <- K

# observed data ----
predictors <- expand.grid(group_1 = group_1, group_2 = group_2)
X_mat <- sparse.model.matrix(~ factor(group_1) + factor(group_2) - 1, data = predictors)

# for sparsity, since here, we're assuming we have only dummies
# creating numeric values for each individual FE
predictors_as_numeric <- cbind(
  as.numeric(factor(predictors[, 1])), as.numeric(factor(predictors[, 2])))

# the regression part of the equation
betas <- matrix(rnorm(n = ncol(X_mat), 0, 2))
linear_predictor <- X_mat %*% betas
```

Simulate a Factorization Machine:

Next, latent factors:

```
# group_1 factors are gammas
gammas <- mvrnorm(
  n = N,
  mu = rep(0, K),
  Sigma = diag(K))

# group 2 factors are deltas
deltas <- mvrnorm(
  n = J,
  mu = rep(0, K),
  Sigma = diag(K))

factor_terms <-
  matrix(NA, nrow = nrow(linear_predictor), ncol = 1)

for (i in 1:nrow(predictors)) {
  g1 <- as.character(predictors[i, 1])
  g1 <- as.numeric(substr(g1, 2, nchar(g1)))

  g2 <- as.character(predictors[i, 2])
  g2 <- as.numeric(substr(g2, 2, nchar(g2)))

  factor_terms[i,] <- matrix(gammas[g1,], nrow = 1) %*%
    matrix(deltas[g2,], ncol = 1)
}

y <- linear_predictor + factor_terms + rnorm(n = nrow(linear_predictor), 0, y_sigma)
```

Factorization Machines in Stan

Use Stan to fit FMs:

► Data:

```
data{
  int<lower = 0> N ; // number of group 1 observations
  int<lower = 0> J ; // number of group 2 observations
  int<lower = 0> K ; // number of latent dimensions
  int X[(N*J), 2] ; // covariate matrix
  vector[(N*J)] y ; // outcome
  real<lower = 0> beta_sigma ; // sd on regression coefficients
  real<lower = 0> y_sigma ; // sd on the outcome, y
}
```

Factorization Machines in Stan

Use Stan to fit FMs:

► Parameters:

```
parameters{
  vector[N] group_1_betas; // non-interacted coefficients
  vector[J] group_2_betas; // non-interacted coefficients
  matrix[N, K] gammas; // individual factors
  matrix[J, K] deltas; // group 2 factors
}

transformed parameters{
  real linear_predictor[(N*J)] ;
  for(i in 1:(N*J)){
    linear_predictor[i] =
      group_1_betas[X[i, 1]] + group_2_betas[X[i, 2]] +
      (gammas[X[i, 1], ] * deltas[ X[i, 2], ]');
  }
}
```


Factorization Machines in Stan

Use Stan to fit FMs:

► Model:

```
model{  
  
  // regression coefficients  
  group_1_betas ~ normal(0, beta_sigma) ;  
  group_2_betas ~ normal(0, beta_sigma) ;  
  
  // latent factors  
  for(n in 1:N){  
    gammas[n, ] ~ normal(rep_vector(0, K), 1) ;  
  }  
  
  for(j in 1:J){  
    deltas[j, ] ~ normal(rep_vector(0, K), 1) ;  
  }  
  
  // outcome  
  y ~ normal(linear_predictor, y_sigma) ;  
}
```

Factorization Machines in Stan

Use Stan to fit FMs:

- ▶ Model Checking

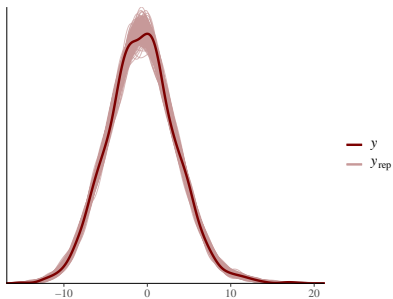
```
generated quantities{  
  real y_pred[(N*J)] ;  
  y_pred = normal_rng(linear_predictor, y_sigma) ;  
}
```

Fitting a Model to Simulated Data

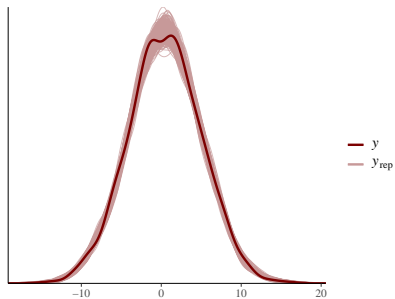
- ▶ Using `simulate_data` in `simulation_function.R`.
- ▶ Two versions:
 - ▶ 100 members of group 1, 20 members of group 2
 - ▶ 20 members of group 1, 100 members of group 2
- ▶ Fit model in `stan`, using NUTS

Fitting a Model to Simulated Data

Model 1 Posterior Predictive



Model 2 Posterior Predictive



Fitting a Model to Simulated Data

TODO: MSE Here!

Now, Let's Get Hierarchical!

Hierarchical Factorization Machines

- ▶ Basic FM implementation assumes all parameters are independent
- ▶ Hierarchical: Sharing parameters across groups
- ▶ Sharing parameters $\$ = \$$ sharing information!

Hierarchical Factorization Machines

$$y_{nj} \sim \mathcal{N}(\mathbf{x}_n \beta_1 + \mathbf{x}_j \beta_2 + \gamma_n \cdot \delta_j^\top, \sigma_y)$$

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta)$$

$$\gamma_n, \sim \mathcal{N}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$$

$$\delta_j, \sim \mathcal{N}(\boldsymbol{\mu}_\delta, \boldsymbol{\Sigma}_\delta)$$

$$\mu_\gamma \sim \mathcal{N}(0, 1)$$

$$\mu_\delta \sim \mathcal{N}(0, 1)$$

$$\boldsymbol{\Sigma}_\gamma = \sigma_\gamma \boldsymbol{\Omega}_\gamma$$

$$\boldsymbol{\Sigma}_\delta = \sigma_\delta \boldsymbol{\Omega}_\delta$$

Hierarchical Factorization Machines in Stan