Nonparametric Bayes tensor factorizations for big data

David Dunson

Department of Statistical Science, Duke University

Funded from NIH R01-ES017240, R01-ES017436 & DARPA N66001-09-C-2082

Motivation

Conditional tensor factorizations

Some properties - heuristic & otherwise

Computation & applications

Generalizations

Motivating setting - high dimensional predictors

- Routine to encounter massive-dimensional prediction & variable selection problems
- We have $y \in \mathcal{Y}$ & $x = (x_1, \ldots, x_p)' \in \mathcal{X}$
- Unreasonable to assume linearity or additivity in motivating applications - e.g., epidemiology, genomics, neurosciences
- Goal: nonparametric approaches that accommodate large p, small n, allow interactions, scale computationally to big p

Gaussian processes with variable selection

- For $\mathcal{Y} = \Re$ & $\mathcal{X} \subset \Re^p$, then one approach lets

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

  where $\mu : \mathcal{X} \to \Re$ is an unknown regression function
- Following Zou et al. (2010) & others,

$$\mu \sim \mathrm{GP}(m, c), \quad c(x, x') = \phi \exp\left\{ -\sum_{j=1}^{p} \alpha_j (x_j - x'_j)^2 \right\},$$

  with mixture priors placed on α_j's
- Zou et al. (2010) show good empirical results
- Bhattacharya, Pati & Dunson (2011) - minimax adaptive rates

Issues & alternatives

- Mean regression & computation challenging
- Difficult computationally beyond conditionally Gaussian homoscedastic case
- Density regression interesting as variance & shape of response distribution often changes with x
- Initial focus: classification from many categorical predictors
- Approach generalizes directly to arbitrary $\mathcal{Y}$ and $\mathcal{X}$.

Classification & conditional probability tensors

- Suppose $Y \in \{1, \ldots, d_0\}$ & $X_j \in \{1, \ldots, d_j\}$,j=1,. . . ,p
- The classification function or conditional probability is

$$\Pr(Y = y | X_1 = x_1, \ldots, X_p = x_p) = P(y | x_1, \ldots, x_p).$$

- This classification function can be structured as a $d_0 \times d_1 \times \cdots \times d_p$ tensor
- Let $\mathcal{P}_{d_1,\ldots,d_p}(d_0)$ denote to set of all possible *conditional probability tensors*
- $P \in \mathcal{P}_{d_1,\ldots,d_p}(d_0)$ implies $P(y | x_1, \ldots, x_p) \geq 0 \ \forall y, x_1, \ldots, x_p$ & $\sum_{y=1}^{d_0} P(y | x_1, \ldots, x_p) = 1$

Tensor factorizations

- P = big tensor & data will be very sparse
- If P was a matrix, we may think of SVD
- We can instead consider a tensor factorization
- Common approach is PARAFAC - sum of rank one tensors
- Tucker factorizations express $d_1 \times \cdots \times d_p$ tensor $A = \{a_{c_1 \cdots c_p}\}$ as

$$a_{c_1 \cdots c_p} = \sum_{h_1=1}^{d_1} \cdots \sum_{h_p=1}^{d_j} g_{h_1 \cdots h_p} \prod_{j=1}^{p} u^{(j)}_{h_j c_j},$$

where $G = \{g_{h_1 \cdots h_p}\}$ is a core tensor,

Our factorization (*with Yun Yang*)

- Our proposed nonparametric model for the conditional probability:

$$P(y|x_1,\ldots,x_p) = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \lambda_{h_1 h_2 \ldots h_p}(y) \prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_j) \qquad (1)$$

- Tucker factorization of the conditional probability P
- To be valid conditional probability, parameters subject to

$$\sum_{c=1}^{d_0} \lambda_{h_1 h_2 \ldots h_p}(c) = 1, \text{ for any } (h_1, h_2, \ldots, h_p),$$

$$\sum_{h=1}^{k_j} \pi_h^{(j)}(x_j) = 1, \text{ for any possible pair of } (j, x_j). \qquad (2)$$

Comments on proposed factorization

- $k_j = 1$ corresponds to exclusion of the jth feature
- By placing prior on k_j, can induce variable selection & learning of dimension of factorization
- Representation is many-to-one and the parameters in the factorization cannot be uniquely identified.
- Does not present a barrier to Bayesian inference - we don't care about the parameters in factorization
- We want to do variable selection, prediction & inferences on predictor effects

Theoretical support

The following Theorem formalizes the flexibility:

Theorem

Every $d_0 \times d_1 \times d_2 \times \cdots \times d_p$ conditional probability tensor $P \in \mathcal{P}_{d_1,\ldots,d_p}(d_0)$ can be decomposed as (1), with $1 \leq k_j \leq d_j$ for $j = 1, \ldots, p$. Furthermore, $\lambda_{h_1 h_2 \ldots h_p}(y)$ and $\pi^{(j)}_{h_j}(x_j)$ can be chosen to be nonnegative and satisfy the constraints (2).

Latent variable representation

- Simplify representation through introducing p latent class indicators $z_1, \ldots, z_p$ for $X_1, \ldots, X_p$
- Conditional independence of Y and $(X_1, \ldots, X_p)$ given $(z_1, \ldots, z_p)$
- The model can be written as

$$\begin{aligned} Y_i | z_{i1}, \ldots, z_{ip} &\sim \text{Mult}\big(\{1, \ldots, d_0\}, \boldsymbol{\lambda}_{z_{i1}, \ldots, z_{ip}}\big), \\ z_{ij} | X_{ij} = x_j &\sim \text{Mult}\big(\{1, \ldots, k_j\}, \pi_1^{(j)}(x_j), \ldots, \pi_{k_j}^{(j)}(x_j)\big), \end{aligned}$$

- Useful computationally & provides some insight into the model

Prior specification & hierarchical model

- Conditional likelihood of response is $(Y_i|z_{i1},\ldots,z_{ip},\Lambda) \sim$

$$\text{Multinomial}\big(\{1,\ldots,d_0\},\boldsymbol{\lambda}_{z_{i1},\ldots,z_{ip}}\big)$$

- Conditional likelihood of latent class variables is

$$(z_{ij}|X_{ij}=x_j,\pi) \sim \text{Multinomial}\big(\{1,\ldots,k_j\},\pi_1^{(j)}(x_j),\ldots,\pi_{k_j}^{(j)}(x_j)\big)$$

- Prior on core tensor $\boldsymbol{\lambda}_{h_1,\ldots,h_p} =$

$$\big(\lambda_{h_1,\ldots,h_p}(1),\ldots,\lambda_{h_1,\ldots,h_p}(d_0)\big) \sim \text{Diri}(1/d_0,\ldots,1/d_0)$$

- Prior on independent rank one components,

$$\big(\pi_1^{(j)}(x_j),\ldots,\pi_{k_j}^{(j)}(x_j)\big) \sim \text{Diri}(1/k_j,\ldots,1/k_j)$$

Prior on predictor inclusion/tensor rank

- For the jth dimension, we choose the simple prior

$$P(k_j = 1) = 1 - \frac{r}{p}, \; P(k_j = k) = \frac{r}{(d_j - 1)p}, k = 2, \ldots, d_j,$$

  d_j=# levels of covariate X_j.
- r=expected # important features, $\bar{r}$=specified maximum number of features
- Effective prior on k_j's is $P(k_1 = l_1, \ldots, k_p = l_p) =$

$$P(k_1 = l_1) \cdots P(k_p = l_p) I_{\{\sharp\{j: l_j > 1\} \leq \bar{r}\}}(l_1, \ldots, l_p),$$

  where $I_A(\cdot)$ is the indicator function for set A.

Properties - Bias-Variance Tradeoff

- Extreme data sparsity - vast majority of combinations of $Y, X_1, \ldots, X_p$ not observed
- Critical to include sparsity assumptions - even if such assumptions do not hold, massively reduces the variance
- Discard predictors having small impact & parameters having small values
- Makes the problem tractable & may lead to good MSE

Illustrative example

- Binary Y & p binary covariates $X_j \in \{-1, 1\}$, $j = 1, \ldots, p$
- The true model can be expressed in the form $[\beta \in (0, 1)]$

$$P(Y = 1 | X_1 = x_1, \ldots, X_p = x_p) = \frac{1}{2} + \frac{\beta}{2^2} x_1 + \cdots + \frac{\beta}{2^{p+1}} x_p.$$

  Effect of X_j decreases exponentially as j increases from 1 to p.
- Natural strategy: estimate $P(Y = 1 | X_1 = x_1, \ldots, X_p = x_p)$ by sample frequencies over 1st k covariates

$$\frac{\sharp\{i : y_i = 1, x_{1i} = x_1, \ldots, x_{ki} = x_k\}}{\sharp\{i : x_{1i} = x_1, \ldots, x_{ki} = x_k\}},$$

  & ignore the remaining $p - k$ covariates.
- Suppose we have $n = 2^l$ $(k \leq l \ll p)$ observations with one in each cell of combinations of $X_1, \ldots, X_l$.

MSE analysis

- Mean square error (MSE) can be expressed as

$$\begin{aligned}\text{MSE} &= \sum_{h_1,\ldots,h_p} E\big\{P(Y=1|X_1=h_1,\ldots,X_p=h_p) - \\ &\qquad \hat{P}(Y=1|X_1=h_1,\ldots,X_k=h_k)\big\}^2 \\ &\triangleq \text{Bias}^2 + \text{Var}.\end{aligned}$$

- The squared bias is

$$\begin{aligned}\text{Bias}^2 &= \sum_{h_1,\ldots,h_p} \big\{P(Y=1|X_1=h_1,\ldots,X_p=h_p) - \\ &\qquad E\hat{P}(Y=1|X_1=h_1,\ldots,X_k=h_k)\big\}^2 \\ &= \beta^2 2^{k+1} \sum_{i=1}^{2^{p-k-1}} \left(\frac{2i-1}{2^{p+1}}\right)^2 = \frac{\beta^2}{3}(2^{p-2k-2} - 2^{-p-2}).\end{aligned}$$

MSE analysis (continued)

- Finally we obtain the variance as

$$\begin{aligned}
\text{Var} &= \sum_{h_1,\ldots,h_p} \text{Var}\hat{P}(Y=1|X_1=h_1,\ldots,X_k=h_k) \\
&= 2^{p-k+1}\sum_{i=1}^{2^{k-1}} \frac{1}{2^l}\left(\frac{1}{2}+\frac{2i-1}{2^{k+1}}\beta\right)\left(\frac{1}{2}-\frac{2i-1}{2^{k+1}}\beta\right) \\
&= \frac{1}{3}\{(3-\beta^2)2^{p+k-l-2}+\beta^2 2^{p-k-l-2}\}.
\end{aligned}$$

- Since there are 2^p cells, the average MSE for each cell equals

$$\frac{1}{3}\{(3-\beta^2)2^{k-l-2}+\beta^2 2^{-k-l-2}+\beta^2 2^{-2k-2}-\beta^2 2^{-2p-2}\}.$$

Implications of MSE analysis

- #predictors p has little impact on selection of k
- $k \leq l$ & so second term small comparing to 1st & 3rd terms
- Average MSE obtains its minimum at $k \approx l/3 = \log_2(n)/3$
- True model not sparse & all the predictors impact conditional probability
- But optimal # predictors only depends on the log sample size

Borrowing of information

- Critical feature of our model is *borrowing* across cells
- Letting $w_{h_1,\ldots,h_p}(x_1,\ldots,x_p) = \prod_j \pi^{(j)}_{h_j}(x_j)$, our model is

$$P(Y = y | X_1 = x_1, \ldots, X_p = x_p) = \sum_{h_1,\ldots,h_p} w_{h_1,\ldots,h_p}(x_1,\ldots,x_p)\lambda_{h_1\ldots h_p}(y),$$

  with $\sum_{h_1,\ldots,h_p} w_{h_1,\ldots,h_p}(x_1,\ldots,x_p) = 1$.
- View $\lambda_{h_1\ldots h_p}(y)$ as frequency of $Y = y$ in cell $X_1 = h_1, \ldots, X_p = h_p$
- We have kernel estimate for borrowing information via weighted avg of cell freqs

Illustrative example

- One covariate $X \in \{1, \ldots, m\}$ with $Y \in \{0, 1\}$ & $P_j = P(Y = 1 | X = j)$
- Naive estimate $\hat{P}_j = k_j / n_j = \sharp\{i : y_i = 1, x_i = j\} / \sharp\{i : x_i = j\}$ = sample freqs
- Alternatively, consider kernel estimate indexed by $0 \leq c \leq 1/(m-1)$

$$\tilde{P}_j = \{1 - (m-1)c\}\hat{P}_j + c \sum_{k \neq j} \hat{P}_k, \; j = 1, \ldots, m.$$

- Use squared error loss to compare these estimators

MSE for illustrative example

- $E\{L(\hat{P}, P)\} = \sum_{j=1}^{m} E(\hat{P}_j - P_j)^2 = \sum_{j=1}^{m} \frac{P_j(1-P_j)}{n_j}$.
- $E\{L(\tilde{P}, P)\} = \sum_{j=1}^{m} E(\tilde{P}_j - P_j)^2$ is fn of c with min at

$$c_0 = \frac{1}{m} \frac{E\{L(\hat{P}, P)\}}{E\{L(\hat{P}, P)\} + \frac{1}{m-1} \sum_{i<j} (P_i - P_j)^2} \in \left(0, \frac{1}{m-1}\right).$$

- When P_j's are similar, estimate $\tilde{P}$ can reduce risk up to only $1/m$ the risk of estimating $\hat{P}$ separately.
- If P_j's are not similar, $\tilde{P}$ can still reduce the risk considerably when the cell counts $\{n_j\}$ are small.

Setting & assumptions

- Data y^n & X^n for n subjects with $p_n \gg n$ (*large p, small n*)
- Assume $d_j = d$ for simplicity in exposition
- Putting true model P_0 in our tensor form, assume

  Assumption A. $\sum_{j=1}^{p_n} \max_{x_j} \sum_{h_j=2}^{d} \pi_{h_j}^{(j)}(x_j) < \infty$.

  This is a near sparsity restriction on P_0.
- Additionally assume true conditional probabilities strictly greater than zero,

  Assumption B. $P_0(y|x) \geq \epsilon_0$ *for any* x, y *for some* $\epsilon_0 > 0$.

Posterior convergence theorem

- $x_1, \ldots, x_n$ independent from unknown G_n on $\{1, \ldots, d\}^{p_n}$
- Let ϵ_n be a sequence with $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$ and $\sum_n \exp(-n\epsilon_n^2) < \infty$.
- Assume the following conditions hold: (i) $\bar{r}_n \log p_n \prec n\epsilon_n^2$, (ii) $\bar{r}_n d^{\bar{r}_n} \log(\bar{r}_n/\epsilon_n) \prec n\epsilon_n^2$, (iii) $r_n/p_n \to 0$ as $n \to \infty$, and (iv) there exists a sequence of models γ_n with size $\bar{r}_n$ such that $\sum_{j \notin \gamma_n} \max_{x_j} \sum_{h_j=2}^{d} \pi_{h_j}^{(j)}(x_j) \prec \epsilon_n^2$.
- Denote $d(P, P_0) = \int \sum_{y=1}^{d_0} \left| P(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p) \right| G_n(dx_1, \ldots, dx_p)$, then

$$\Pi_n\left\{P : d(P, P_0) \geq M\epsilon_n | y^n, X^n\right\} \to 0 \ a.s. P_0^n.$$

Implications of theorem

- Posterior convergence rate can be very close to $n^{-1/2}$ for appropriate hyperparameter choices.
- For any $\alpha \in (0,1)$, $\epsilon_n = n^{-(1-\alpha)/2} \log n$ satisfies conditions
  - $r_n \prec \bar{r}_n \prec \log n$ (*# important predictors scales w/ log n*)
  - $p_n \prec \exp(n^\alpha)$ (*# candidate predictors exponential in n*)
  - There exists a sequence of models γ_n with size $\bar{r}_n$ such that

$$\sum_{j \notin \gamma_n} \max_{x_j} \sum_{h_j=2}^{d} \pi_{h_j}^{(j)}(x_j) \prec n^{\alpha-1} \log^2 n.$$

- Use $r_n = \log_d(n), \bar{r}_n = 2r_n$ as default values for the prior in applications.

Posterior computation

- Conditionally on $\{k_j\}$ simple Gibbs sampler - Dirichlet & multinomial conditionals
- # components to update $\prod_{j=1}^{p} k_j$ - can blow up with p but all but small number of $k_j = 1$
- To update $\{k_j\}$ we can use RJMCMC - current vs doesn't scale very well computationally
- Two stage algorithm: (i) SSVS to estimate k_j - acceptance probs use approximated conditional marginal likelihoods; (ii) conditionally on $\{\hat{k}_j\}$ run Gibbs
- Scales efficiently & excellent performance in cases we have considered

Simulation study

- $N = 2,000$ instances, $p = 600$ covariates $X_j \in \{1, \ldots, 4\}$ & binary response Y
- True model: 3 important predictors X_9, X_{11} and X_{13}
- Generate $P(Y = 1 | X_9 = x_9, X_{11} = x_{11}, X_{13} = x_{13})$ independently for each combination of (x_9, x_{11}, x_{13}).
- To obtain optimal misclassification rate $\sim 15\%$, generated

$$f(U) = U^2 / \{U^2 + (1 - U)^2\}, \quad U \sim \text{Unif}(0, 1).$$

- n training samples & $N - n$ testing
- Training - $n \in \{200, 400, 600, 800\}$, with 10 random training-test splits. Apply our approach to each split.

Simulation results - misclassification rate

Table: Testing Results for Synthetic Data Example. RF: random forests; TF: Our tensor factorization model.

training size	200	400	600	800
aMSE of TF	0.144	0.042	0.024	0.010
Misclassification Rate of TF	0.503	0.288	0.189	0.168
Misclassification Rate of RF	0.496	0.482	0.471	0.472

$$\text{aMSE} = \frac{1}{4^p} \sum_{x_1,\ldots,x_p} \left\{ P(Y = 1|x_1, \ldots, x_p) - \hat{P}(Y = 1|x_1, \ldots, x_p) \right\}^2,$$

Simulation results - variable selection performance

Table: Columns 2-4 = inclusion probs of 9th,11th,13th predictors.
Col 5 = max inclusion prob across remaining predictors.
Col 6 = average inclusion probability across the remaining predictors.
Quantities are averages over 10 trials.

training size	9	11	13	Max	Average
200	0.092	0.041	0.063	0.161	0.002
400	0.816	0.820	0.808	0.013	0.000
600	1.000	1.000	1.000	0.000	0.000
800	1.000	1.000	1.000	0.000	0.000

Application data sets

1. Promoter gene sequences: A, C, G, T nucleotides at $p = 57$ positions for $N = 106$ sequences & binary response (promoter or not). 5-fold CV - $n = 85$ training & $N - n = 21$ test samples in each split.
2. Splice-junction gene sequences: A, C, G, T nucleotides at $p = 60$ positions for $N = 3,175$ sequences. response classes: exon/intron boundary (EI), intron/exon boundary (IE) or neither (N). Test - $n \in \{200, 2540\}$.
3. Single Proton Emission Computed Tomography (SPECT): cardiac patients normal/abnormal. 267 SPECT images & 22 binary features. Previously divided - $n = 80$ & $N - n = 187$.

Results

Table: RF: random forests, NN: neural networks, SVM: support vector machine, BART: Bayesian additive regression trees, TF: Our tensor factorization model. Misclassification rates are displayed.

Data	CART	RF	NN	LASSO	SVM	BART	TF
Promoter (n=85)	0.236	**0.066**	0.170	0.075	0.151	0.113	**0.066**
Splice (n=200)	0.161	0.122	0.226	0.141	0.286	-	**0.112**
Splice (n=2540)	0.059	**0.046**	0.165	0.123	0.059	-	0.058
SPECT (n=80)	0.312	0.235	0.278	0.277	0.246	0.225	**0.198**

- At worst comparable classification performance with RF best of competitors
- Particularly good relative performance as n decreases & p increases

Variable selection - interpretability & parsimony

- Additional advantages in terms of variable selection
- In promoter data, selected nucleotides at 15th, 16th, 17th, and 39th positions
- In splice data, 28th, 29th, 30th, 31st, 32nd and 35th positions are selected.
- In SPECT data, 11st, 13rd and 16th predictors are selected.
- Each case obtained excellent classification performance based on a small subset of the predictors.

Generalization - conditional distribution modeling

- Generalization: conditional distribution estimation

$$f(y|x) = \sum_{h=1}^{k} \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \pi_{hh_1 \cdots h_p}(x) \mathcal{K}(y; \theta_{hh_1 \cdots h_p}),$$

- $\{\pi_{hh_1 \cdots h_p}(x)\}$ = core probability tensor of predictor-dependent weights on a multiway array of kernels
- Motivated by above conditional tensor factorization for classification, let

$$\pi_{hh_1 \cdots h_p} = \pi_h \prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_j).$$

Linear Tucker density regression

- Letting $x \in \mathcal{X} = [0,1]^p$ and $\psi_j \in [0,1]$, choose

$$\pi_1^{(j)}(x_j) = 1 - x_j\psi_j, \quad \pi_2^{(j)}(x_j) = x_j\psi_j, \quad k_j = 2,\ j = 1,\ldots,p,$$

- Model linearly interpolates but otherwise is extremely flexible
- In simple case in which $p = 1$ & Gaussian kernel, we have

$$f(y|x) = \sum_{h=1}^{k} \pi_h \{(1 - x\psi)N(y; \mu_{h1}, \tau_{h1}^{-1}) + x\psi N(y; \mu_{h2}, \tau_{h2}^{-1})\},$$

- Induces the linear mean regression model

$$E(y|x) = \left\{ \sum_{h=1}^{k} \pi_h \mu_{h1} \right\} + \left\{ \sum_{h=1}^{k} \pi_h \psi(\mu_{h2} - \mu_{h1}) \right\} x = \beta_0 + \beta_1 x,$$

Linear Tucker density regression - comments

- Different quantiles of $f(y|x)$ change linearly with x but with slopes that vary
- If $k = 1$, and $\tau_{h1} = \tau_{h2} = \tau_h$, obtain simple normal linear regression
- Density changes linearly - $f(y|x = 0) = \sum_h \pi_h N(y; \mu_{h1}, \tau_{h1}^{-1})$ to $f(y|x = 1) = \sum_h \pi_h N(y; \mu_{h2}, \tau_{h2}^{-1})$ as x increases
- As p increases, still interpolate linearly but accommodate interactions
- Posterior computation single stage Gibbs sampler (*multinomial, Dirichlet, normal-gamma, Bernoulli, beta*)

Joint tensor factorizations

- Focused in this talk on conditional Tucker factorizations
- Can also use probabilistic tensor factorizations for *joint* modeling
- Very useful for huge sparse contingency table analysis
- Same ideas provide type of multivariate generalization of current Bayes discrete mixtures
- Instead of a single cluster index, multiple dependent cluster indices underlying each type of data

References

- Banerjee, A., Murray, J. & Dunson, D.B. (2012). Nonparametric Bayes infinite tensor factorization priors.
- Bhattacharya, A. and Dunson, D.B. (2012). Simplex factor models for multivariate unordered categorical data. *JASA*, 107, 362-377.
- Bhattacharya, A. and Dunson, D.B. (2012). Nonparametric Bayes testing of associations in high-dimensional categorical data. *almost done!*
- Dunson, D.B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *JASA*, 104, 1042-1051.
- Ghosh, A. and Dunson, D.B. (2012). Conditional distribution from high-dimensional interacting predictors. *in progress*.
- Yang, Y. and Dunson, D.B. (2012). Bayesian conditional tensor factorizations for high-dimensional classification. *arXiv*.