BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1562–1568
doi:10.1093/bioinformatics/btu040

Genetics and population analysis

Advance Access publication February 5, 2014

Learning phenotype densities conditional on many interacting predictors

David C. Kessler[1,*], Jack A. Taylor[2] and David B. Dunson[3]

[1]Advanced Analytics Division, SAS Institute Inc., Cary, NC 27513, [2]Molecular and Genetic Epidemiology Section, Epidemiology Branch and Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709 and [3]Department of Statistical Science, Duke University, Durham, NC 27708

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Estimating a phenotype distribution conditional on a set of discrete-valued predictors is a commonly encountered task. For example, interest may be in how the density of a quantitative trait varies with single nucleotide polymorphisms and patient characteristics. The subset of important predictors is not usually known in advance. This becomes more challenging with a high-dimensional predictor set when there is the possibility of interaction.

Results: We demonstrate a novel non-parametric Bayes method based on a tensor factorization of predictor-dependent weights for Gaussian kernels. The method uses multistage predictor selection for dimension reduction, providing succinct models for the phenotype distribution. The resulting conditional density morphs flexibly with the selected predictors. In a simulation study and an application to molecular epidemiology data, we demonstrate advantages over commonly used methods.

Availability and implementation: MATLAB code available at https://googledrive.com/host/0Bw6KlFB-k4lOOWQ0dFJtSVZxNE0/ktdctf.html

Contact: dave.kessler@gmail.com

Received on July 19, 2013; revised on December 29, 2013; accepted on January 17, 2014

1 INTRODUCTION

Many areas of research are concerned with learning the distribution of a response conditional on numerous categorical (discrete) predictors. The important predictors for characterization of this distribution are not usually known in advance, and there may be hundreds or thousands of candidates. Methods that attempt to accommodate interactions among these predictors become mired in the enormous model space. For example, in a moderate-dimensional case involving $p = 40$ categorical predictors, each with $d_j = 4$ possible realizations, considering all possible levels of interaction leads to a space of $4^{40} \approx 10^{24}$ possible models. Parallelization and technical tricks may work for smaller examples, but data sparsity and the sheer volume of models force us to consider different approaches. The conditional density may vary in more than just location; Chung and Dunson (2009) illustrated this in an application to the conditional density of blood glucose levels given insulin sensitivity and age. In the work that follows, we present a novel non-parametric Bayes (NPB) approach to learning conditional densities that makes use of a conditional tensor factorization to characterize the conditional distribution given the predictor set, allowing for complex interactions between the predictors. The particular form assumed for the conditional density gives rise to an attractive predictor selection procedure, providing support for distinct predictor selection steps. This addresses the challenges of high-dimensional data and produces conditional density estimates that allow assessment of tail risks and other complex quantities.

*To whom correspondence should be addressed.

2 APPROACH

The primary goal of our work is to model the conditional density $f(y|\boldsymbol{x})$, where the form of this density for the response y changes flexibly with the predictor vector $\boldsymbol{x}$. There is a large body of work devoted to this idea of density regression in settings involving $\boldsymbol{x}$ of dimension $p \leq 30$, and such models have provided many options for that situation. We wish to develop techniques for problems involving much larger p, and ideally to scenarios where $p > 1000$. We want to provide a method that performs variable selection, assesses the probability of a predictors inclusion in the model and provides easily interpretable estimates of the impact of different predictors. This problem has been addressed with variations on the finite mixture model,

$$f(y) = \sum_{h=1}^{K} \pi_h \mathcal{K}(y; \theta_h) \quad (1)$$

This is the basic form of the hierarchical mixture of experts model [HME, Jordan and Jacobs (1994)]. In this representation, K represents the number of contributing parametric kernels $\mathcal{K}(;\theta_h)$ distinguished by parameters θ_h. The π_h provides the weights in this convex combination of kernels, where $\sum_{h=1}^{K} \pi_h = 1$ and $(\pi_1, \ldots, \pi_K) \in \mathcal{S}_{K-1}$, the $K-1$ probability simplex. The most straightforward forms rely on a pre-specified K and include the predictors $\boldsymbol{x}$ in a linear model for the mean. HME methods in the frequentist literature have often relied on expectation maximization (EM) (Dempster *et al.*, 1977) techniques, which can suffer from overfitting (Bishop and Svensén, 2003). EM approaches in the Bayesian literature seek to avoid this; Waterhouse *et al.* (1996) used EM to find maximum a posteriori estimates using the inherent Bayesian penalty against complexity to regulate those estimates. In addition, the Bayesian framework allows the quantification of uncertainty about the parameters in the model.

NPB methods, such as the Dirichlet Process, prompted techniques like that in Muller *et al*. (1996), which induced flexible conditional regression through joint modeling of the response and predictors. Subsequent methods included the predictors in π_h and/or θ_h via Dependent Dirichlet Process (DDP) mixtures. De Iorio *et al*. (2004) proposed an ANOVA DDP model with fixed weights $\{\pi_h\}$ that used a small number of categorical predictors to index random distributions for the response. Griffin and Steel (2006) developed an ordered DDP, where the predictor vectors were mapped to specific permutations of the weights $\{\pi_h\}$, yielding different density estimates for different predictor vectors. Reich and Fuentes (2007) and Dunson and Park (2008) used the kernel stick-breaking process to allow predictors to influence the weights. Chung and Dunson (2009) presented a further alternative in the probit stick-breaking process, which uses a probit transform of a real-valued function of the predictors to incorporate them into the weights. Methods that use joint modeling of response and predictors (Shahbaba and Neal, 2009; Hannah *et al*., 2011; Dunson and Xing, 2009) are popular and can work well under many circumstances, but estimation of the marginal distribution of the predictors is a burden. Methods not relying on discrete mixtures also exist; Tokdar *et al*. (2010) developed a technique based on logistic Gaussian processes. Jara and Hanson (2011) presented an approach using mixtures of transformed Gaussian processes.

These and other methods of Bayesian density regression have proven successful, but as datasets have grown in size and complexity, these approaches encounter difficulties. This is even more daunting when we consider interactions of discretely valued predictors because we must consider the factorial combinations of those levels.

The associated challenges of variable selection and dimensionality reduction have been explored in Bayesian density regression. Dimensionality reduction has a goal similar to that of variable selection, that of finding a minimal set of predictors that account for variation in the response. The logistic Gaussian process approach of Tokdar *et al*. (2010) includes a subspace projection method to reduce the dimension of the predictor space. Reich *et al*. (2011) developed a technique for Bayesian sufficient dimensionality reduction based on a prior for a central subspace. Although all of these approaches have demonstrated their utility, they do not scale easily beyond $p = 30$ predictors.

There are also techniques like the random forest (Breiman, 2001) that aim to find parsimonious models for density estimation involving a large number of predictors. One disadvantage to this type of 'black box' method is in interpreting the impact of specific predictors on the response. Bayesian additive regression trees (BART) (Chipman *et al*., 2006, 2010) focus on modeling the conditional mean and assume a common residual distribution. As previously noted, there are many questions that require learning about more than just the conditional mean of the response. Another flexible approach is the Bayes network (BN), which considers the predictors and the response on equal footing to develop a parsimonious network linking all variables (Pearl, 1988; Cowell *et al*., 1999; Lauritzen, 1992). The conditional distribution of the response given the predictors can be derived from such a model, using developed BN techniques for mixed continuous and discrete data (Lauritzen, 1992; Moral *et al*., 2001; Langseth *et al*., 2012). A BN does estimate a joint density for all of the predictors; the effort to estimate this very high-dimensional nuisance parameter is unattractive, if the conditional density is of primary interest.

We propose an approach based on a conditional tensor factorization (CTF) for the mixing weights. As in the DDP and certain of the kernel stick-breaking methods, the predictors influence the mixing weights for this CTF model. The conditional tensor factorization facilitates borrowing of information across different profiles in a flexible representation of the unknown density. We focus our attention on situations involving continuous responses and categorical predictors.

3 METHODS

We consider a univariate response y and a vector of p categorical predictors $\boldsymbol{x} = (x_1, \ldots, x_p)$, where the j^{th} predictor x_j can take values $1, \ldots, d_j$. We would like a model that can flexibly accommodate conditional densities that change in complex ways with changes in the predictor vector. In addition, we must consider situations where $p \gg n$; there may be no exemplars for certain predictor vectors. To address this, we propose a Tucker-style factorization with the following general model for the conditional density $f(y|x)$:

$$f(y|\boldsymbol{x}) = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \pi_{h_1,\ldots,h_p}(\boldsymbol{x}) \lambda(y; \theta_{h_1,\ldots,h_p})$$

$$\text{where} \quad \pi_{h_1,\ldots,h_p}(\boldsymbol{x}) = \prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_j). \tag{2}$$

This form uses the maps $\pi^{(j)}, j = 1, \ldots, p$ to associate the predictor vector $\boldsymbol{x}$ with a separate weight for each combination of the latent identifiers $h_1, \ldots, h_p$ and thus with each of the $k_1 \times \cdots \times k_p$ kernels in the representation. The x_j^{th} row of $\pi^{(j)}$ is a vector of weights, one for each $h_j = 1, \ldots, k_j$. These weights $\pi_1^{(j)}(x_j), \ldots, \pi_{k_j}^{(j)}(x_j)$ are all in [0,1] and $\sum_{h_j=1}^{k_j} \pi_{h_j}^{(j)}(x_j) = 1$. The number of latent predictors p, is the same as the number of observed predictors, but the form of the $\pi_{h_j}^{(j)}$ may mean that different predictor vectors $\boldsymbol{x}$ result in the same sets of weights $\pi_{1,\ldots,1}(\boldsymbol{x}), \ldots, \pi_{k_1,\ldots,k_p}(\boldsymbol{x})$. This provides the mechanism for dimension reduction that we will develop. An important distinction from the HME is in the treatment of the weights $\pi_{h_1,\ldots,h_p}(\boldsymbol{x})$ as a tensor factorization and the use of kernels $\lambda(y; \theta_{h_1,\ldots,h_p})$, which do not depend on the predictor vector $\boldsymbol{x}$. This is similar in spirit to the classification approach proposed by Yang and Dunson, 2012, but we address the problem of estimating an infinite-dimensional conditional density rather than the finite-dimensional problem of a categorical response distribution. In addition, we make distinct improvements in predictor selection to allow the approach to scale to larger numbers of candidate predictors.

Tucker decompositions (Tucker, 1966) and other kinds of decompositions have appeared in the machine learning literature before. Xu *et al*. (2012) developed an 'infinite' Tucker decomposition making use of latent Gaussian processes rather than explicit treatment of tensors and matrices; in comparison, the proposed method uses the Tucker decomposition to characterize the mapping of predictors into weights. Other factorizations have been used for similar problems; Hoff (2011) presented a reduced-rank approach for table data, but this approach focused on the development of estimates for the mean of a continuous response. Chu and Ghahramani (2009) derive an approach for partially observed multiway data based on a Tucker decomposition; their objective is to learn

about the latent factors driving observations rather than the characterization of the response distribution or variable selection.

The collection across $j = 1, \ldots, p$ forms a 'soft' clustering from the $d_1 \times \cdots \times d_p$ possible realizations of the $\boldsymbol{x}$ vector to each of the $M = k_1 \times \cdots \times k_p$ possible latent vectors. This means that a predictor vector $\boldsymbol{x}$ is not exclusively associated with one of the M kernels, but has a weight for each kernel determined by the product in (2). This allows each observation to contribute some information about the influence of each of the p sites, and thus allows borrowing of information across different combinations of $h_1, \ldots, h_p$. In settings of extreme sparsity, where most of the possible predictor vectors are not represented, this is an attractive property. This uses many fewer parameters than a full factorial representation and is still flexible enough to represent complex conditional distributions. Finally, we assume normal kernels:

$$f(y_i|\boldsymbol{x}_i) = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \left\{ N(y_i; \mu_{h_1,\ldots,h_p}, \tau^{-1}_{h_1,\ldots,h_p}) \times \prod_{j=1}^{p} \pi^{(j)}_{h_j}(x_{ij}) \right\} \quad (3)$$

This resembles other mixture-based approaches to density estimation as originally specified in (1), but the proposed model for the weights provides the desired support for sparsity and information borrowing as previously discussed. In addition, the kernel-specific means $\mu_{h_1,\ldots,h_p}$ and precisions $\tau_{h_1,\ldots,h_p}$ are not functions of the predictor vector. Figure 1 shows a conditional dependence graph for the model parameters and the observed data.

3.1 Predictor selection

The first task in learning the conditional distribution is to identify those predictors that provide the most information about the response; the second task is to learn the form of the conditional distribution given this set of informative predictors. The $k_1, \ldots, k_p$ parameters indicate the number of latent levels for each predictor. Because each k_j can take on the values $1, \ldots, d_j$, the possible combinations of different values for $k_1, \ldots, k_p$ can be immense, and including these as parameters in an Markov chain Monte Carlo (MCMC) sampler is not an attractive option.

In the notation of (3), predictors exclusion is equivalent to identifying those sites j such that $k_j = 1$. Consequently, predictor vectors that differ only at the j^{th} position will have the same conditional density, and the j^{th} predictor can be excluded from the model. To identify those j such that $k_j = 1$, we use a predictor selection step based on a special form of the $\pi^{(j)}$. For each $j = 1, \ldots, p$ and each $x_j = 1, \ldots, d_j$, we specify the $\pi^{(j)}$ so that $\pi^{(j)}_{h_j}(x_j) = 1$ for exactly one h_j and $\pi^{(j)}_{h_k}(x_j) = 0$ for all $h_k \neq h_j$. This form for the $\pi^{(j)}$ associates each predictor vector $\boldsymbol{x}$ with exactly one of the $M = k_1 \times \cdots \times k_p$ kernels by giving that particular kernel a weight of one. That is, if the set of maps $\pi^{(j)}, j = 1, \ldots, p$ is such that $\pi^{(1)}_{h_1}(x_{i1}) = 1, \ldots, \pi^{(p)}_{h_p}(x_{ip}) = 1$ for values $h_1, \ldots, h_p$, then only the kernel indexed by $h_1, \ldots, h_p$ will have non-zero weight. For computational convenience, we use conjugate priors and make the simplifying assumption that the prior precision of each kernel mean $\mu_{h_1,\ldots,h_p}$ is the same as the kernel precision $\tau_{h_1,\ldots,h_p}$ for each $h_1, \ldots, h_p$, so that

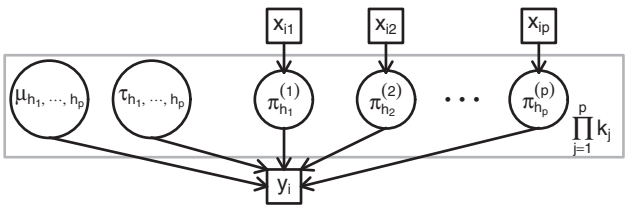

Fig. 1. Conditional dependence graph showing the relationship between the model parameters and the observed data

$\mu_{h_1,\ldots,h_p}|\tau_{h_1,\ldots,h_p} \sim N(0, \tau^{-1}_{h_1,\ldots,h_p})$ and $\tau_{h_1,\ldots,h_p} \sim \text{Gamma}(\delta_t/2, \gamma_t/2)$. Because the proposed form for $\pi^{(1)}, \ldots, \pi^{(p)}$ maps each predictor vector to exactly one of the M groups, we can collect the observations that map to each of the M groups and compute a marginal likelihood for each group. Given the prior structure, the simplifying assumptions and the clusterings defined by the $\pi^{(1)}, \ldots, \pi^{(p)}$, the log marginal likelihood for the m^{th} group is

$$\begin{aligned} &\frac{N_m}{2} log(\pi) - \frac{1}{2} log(N_m + 1) \\ &\quad + log\ \Gamma\left(\frac{N_m + \delta_t}{2}\right) - log\ \Gamma\left(\frac{\delta_t}{2}\right) + \frac{\delta_t}{2}\ log(\gamma_t) \\ &\quad - \frac{1}{2}(N_m + \delta_t)\ log\left\{ Y_m^T Y_m - \frac{(Y_m^T J_{N_m})^2}{N_m + 1} + \gamma_t \right\}, \end{aligned}$$

where Y_m is the vector of responses, N_m is the number of observations in group m and J_{N_m} is a $N_m \times 1$ vector of 1's. The sum of these M approximated log-marginal likelihoods gives a score for the particular levels of $k_1, \ldots, k_p$ and the particular $\pi^{(1)}, \ldots, \pi^{(p)}$. Using these scores for different levels of $k_1, \ldots, k_p$ and different hard-clustering forms of $\pi^{(1)}, \ldots, \pi^{(p)}$, we can find those predictors with influence on the conditional density.

It is not generally feasible to evaluate every possible set of $k_1, \ldots, k_p$, even for moderately sized problems. Instead, we begin with the null model, where $k_1 = k_2 = \ldots = k_p = 1$ and propose random changes to the different k_j and the associated $\pi^{(j)}$. The randomly proposed changes are of two types: 'split' and 'merge'. A 'split' change at position j means changing the $\pi^{(j)}$ map so that the distinct x_{ij} map to more levels. For example, assume that the j^{th} position has three observed levels ($d_j = 3$) and the current form of $\pi^{(j)}$ is such that all three observed levels of x_j are mapped to the same level. In this case, $k_j = 1$ and $\pi^{(j)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. One possible 'split' move would propose $\pi^{(j)}_*$ so that $x_j = 2$ maps to the second latent level, so that $k_j = 2$ and $\pi^{(j)}_* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$. Conversely, a 'merge' move will decrease the number of mapped levels by one; using the definitions above, one such merge move would be to replace $\pi^{(j)}_*$ with $\pi^{(j)}$. If site j already has $k_j = d_j$, then only merge moves are considered. Likewise, if site j already has $k_j = 1$, then only split moves are considered. We use a Metropolis step to accept or reject the proposed change; the stochastic search proceeds as:

(i) Set $n_j = 0, k_j = 1; j = 1, \ldots, p$; set $\pi^{(j)} = J_{d_j}, j = 1, \ldots, p$; compute the marginal likelihood ML^c.
(ii) For $j = 1, \ldots, p$, draw from all possible split and merge moves with equal probability. For a split, propose $k_j^* = k_j + 1$; for a merge, propose $k_j^* = k_j - 1$.
(iii) Compute ML^* for the proposed configuration; accept the move with probability $1 \wedge \frac{ML^*}{ML^c}$. If the new configuration is accepted, set $k_j = k_j^*$ and $ML^c = ML^*$; if $k_j > 1$, set $n_j = n_j + 1$.
(iv) After T iterations of steps 2-3, compute inclusion probabilities $p_j = \frac{n_j}{T}$ for $j = 1, \ldots, p$.
(v) Retain those predictors such that $p_j > \alpha$; using $\alpha = 0.5$ is equivalent to choosing the median probability model.

This stochastic search is similar to George and McCulloch (1997). The approach we propose here is simple and appealing, but in our initial simulation studies we noticed a tendency for this search to choose overly complex models. Model selection was sensitive to the order in which the predictors were considered. When the important features were considered after many unimportant factors, randomly induced associations in the data and stochastic variation in the search led to complex models that were not improved by addition of the important predictors.

To combat this tendency, we introduced a preliminary predictor identification step that considers each of the predictors in isolation. We can represent the entire stochastic search on the j^{th} predictor with a $d_j \times d_j$ discrete-time Markov transition matrix derived from the acceptance and move probabilities defined above. We can then compute the long-run proportion of time that the chain spends in states such that $k_j > 1$. This can be done in an embarrassingly parallel fashion, and the computation of each p_j proceeds quickly. For the simulation case, where $d_j = 4$ for all j, computation of each p_j took ~0.3 s. At the conclusion of this single-predictor search step, we arrange the predictors in descending order of these p_j, retaining only those predictors such that $p_j > \alpha$, and proceed with the full stochastic search to identify a final predictor set.

3.2 Estimation after predictor selection

To estimate the parameters in the model using the selected predictors, we introduce a prior precision $\tau_0 \sim \text{Gamma}(\delta_0/2, \gamma_0/2)$ for each kernel mean $\mu_{h_1,\ldots,h_p} \sim N(0, \tau_0)$, a prior for each kernel precision $\tau_{h_1,\ldots,h_p} \sim \text{Gamma}(\delta_t/2, \gamma_t/2)$ and separate Dirichlet priors for each weight vector $\pi^{(j)}(x_j) \sim \text{Diri}(\frac{1}{k_j}, \ldots, \frac{1}{k_j})$.

To facilitate computation, we augment the model proposed in (3) with classification vectors $\boldsymbol{z}_i$ that associates the i^{th} observation with exactly one kernel and gives a complete-data likelihood that can be expressed as a product:

$$\prod_{i=1}^{N}\prod_{h_1=1}^{k_1}\cdots\prod_{h_p=1}^{k_p}\left\{N\left(y_i;\mu_{h_1,\ldots,h_p},\tau^{-1}_{h_1,\ldots,h_p}\right)\times\prod_{j=1}^{p}\pi^{(j)}_{h_j}(x_{ij})\right\}^{1[\boldsymbol{z}_i=(h_1,\ldots,h_p)]} \quad (4)$$

The full conditional distributions are

(1) $\mu_{h_1,\ldots,h_p}|\cdots \sim N(\mu^*_{h_1,\ldots,h_p}, (\tau^*_{h_1,\ldots,h_p})^{-1})$, where:
$\tau^*_{h_1,\ldots,h_p} = \tau_0 + \tau_{h_1,\ldots,h_p}\sum_{i=1}^{N} 1[\boldsymbol{z}_i = (h_1,\cdots,h_p)]$
$\mu^*_{h_1,\ldots,h_p} = \{\tau_{h_1,\ldots,h_p}\sum_{i=1}^{N} y_i 1[\boldsymbol{z}_i = (h_1,\cdots,h_p)]\}/\tau^*_{h_1,\ldots,h_p}$

(2) $\tau_{h_1,\ldots,h_p}|\cdots \sim \text{Gamma}(\delta^*/2, \gamma^*/2)$, where:
$\delta^* = \delta_t + \sum_{i=1}^{N} 1[\boldsymbol{z}_i = (h_1,\ldots,h_p)]$
$\gamma^* = \gamma_t + \sum_{i=1}^{N} 1[\boldsymbol{z}_i = (h_1,\ldots,h_p)](y_i - \mu_{h_1,\ldots,h_p})^2$

(3) $\tau_0|\cdots \sim \text{Gamma}([\delta_0 + M]/2, [\gamma_0 + \{\sum_{h_1=1}^{k_1}\cdots\sum_{h_p=1}^{k_p}\mu^2_{h_1,\ldots,h_p}\}]/2)$

(4) $(\pi^{(j)}_1(\ell),\ldots,\pi^{(j)}_{k_j}(\ell))|\cdots \sim \text{Diri}(\frac{1}{k_j} + \sum_{i=1}^{N} 1[x_{ij}=\ell]\ 1[z_{ij}=1],\ldots,$
$\frac{1}{k_j} + \sum_{i=1}^{N} 1[x_{ij}=\ell]\ 1[z_{ij}=k_j])$ for $\ell = 1,\ldots,d_j$ and $j = 1,\ldots,p$

(5) $\Pr[\boldsymbol{z}_i = \boldsymbol{z}^*_{jm} \equiv (h_1,\ldots,h_{j-1},m,h_{j+1},\ldots,h_p)]|\cdots \propto \phi[(y_i - \mu_{z^*_{jm}})\sqrt{\tau_{z^*_{jm}}}] \times \pi^{(j)}_m(x_{ij})$ for $m = 1,\ldots,k_j$ within each $j = 1,\ldots,p$; $\phi(\cdot)$ indicates the standard normal density.

The updates for the $\mu_{h_1,\ldots,h_p}$, $\tau_{h_1,\ldots,h_p}$ and $\pi^{(j)}$ can be done blockwise, and the $\boldsymbol{z}_i$ can be updated blockwise at each position j. Using the final predictor set and the full conditionals, we produce a posterior sample for the model parameters. This posterior sample allows us to compute conditional densities and credible intervals around those estimates for various combinations of the predictors.

4 DISCUSSION

4.1 Simulation study

To assess the variable selection and prediction performance of the CTF, we conducted a simulation study, varying the number of training observations $N \in \{300, 500, 1000, 1500\}$ and using a consistent ground truth to produce simulated datasets with total number of predictors $p = 1000$. In each case, the true model was based on three predictors at positions 30, 201 and 801, each with $d_j = 4$ levels and including three-way interactions among these predictors. The resulting marginal density is an equally weighted mixture of 64 Gaussians with different means and the same residual precision τ. In other words, an observation with $(x_{i,30}, x_{i,201}, x_{i,801}) = (3, 2, 1)$ is drawn from $N(\mu_{3,2,1}, \tau^{-1})$, and so forth for each of the 64 distinct predictor vectors.

For each of 20 training sets, we produced selected predictor sets and posterior samples. We then made predictions for 20 validation sets drawn from the same underlying true distribution. As competitor methods, we used random forests (RF) and quantile regression random forests (QRF) (Meinshausen, 2006); these are implemented in the `randomForest` and `quantregForest` packages in `R`. BART, as implemented in the `BayesTree` package, was unable to run to completion on any of the training sets, though we were able to use BART with the real data example in Section 4.2. RF and QRF include predictor selection directly, and QRF directly addresses the idea of coverage proportion. BART is another MCMC-based approach, but it does not directly address variable selection, allowing us to investigate the impact of the large predictor space. The implicit cost in estimating the joint distribution of predictors and response made Bayes networks unattractive.

We computed mean square prediction error (MSPE) as the average squared difference between the response value predicted by the model for a predictor vector from the validation set and the actual response value for that observation. We defined coverage proportion as the proportion of times that the 95% prediction interval for an observation in the validation set included the actual response value, averaged over the intervals for each posterior sample. When comparing performance with that of the competitors, we attempted to give those competitors whatever advantages we could provide. In the case of RF, this meant that we did two passes over the training data. The first pass identified important variables using the `importance` method in the `randomForest` package. We used the 'mean decrease in accuracy' style of importance; this measurement is derived from the impact of permuting out-of-bag data for each tree in the forest. We then fed those variables identified as important as a preselected set into a second run of RF. This generally improved the MSPE performance of RF. An analogous method was not available for QRF, so we could not treat that method in the same manner. In each of the 20 cases for $p = 1000$ and training $N = 500$, the CTF outperformed RF on mean square prediction error and showed comparable 95% coverage proportions to those derived from QRF; this is summarized in Figure 2. The CTF and RF showed comparable accuracy in identifying important predictors, but RF tended to include many unimportant predictors. In contrast, the CTF produced no false-positive results, identifying the correct subset of predictors in each case. This performance is particularly attractive given the large number of possible interactions in the original predictor set. Both RF and QRF may have suffered because of the strong interactions present in these simulated data.

4.2 Molecular epidemiology application

We also consider an application to an epidemiology dataset, comparing CTF performance with that of the same competitor methods (RF, QRF and BART). The dataset concerns DNA damage to instances of different cell lines when exposed to environmental chemicals. The exposure types are hydrogen peroxide (H_2O_2) and methyl methane sulfonate (MMS), and the

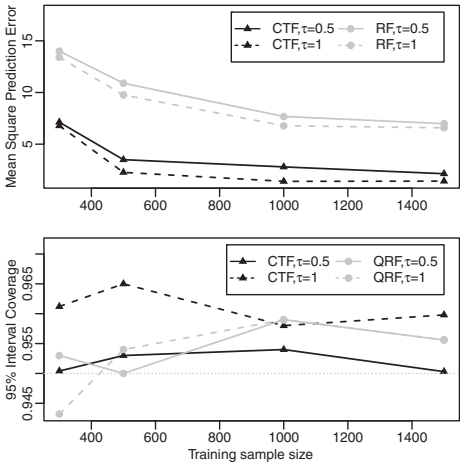

Fig. 2. Simulation study results, comparing CTF with random forests and quantile regression random forests

remainder of the predictor set is genotype information on 49 428 single nucleotide polymorphisms (SNPs). Rodriguez *et al*. (2009) provide extensive details on the original experiments. One hundred separate instances of each of 90 cell lines were exposed to each chemical and examined at each of three time points (before treatment, immediately after treatment and a longer time after treatment). The nature of the measurement is destructive; at the desired time interval, comet assay was performed on each cell and the Olive tail moment (OTM) (Olive *et al*., 1991) was recorded; this assesses the amount of DNA damage in the cell, with higher measurements indicating more damage. The cells from each line are genetically identical, but the resulting distribution of OTM has a different shape for each cell line. In addition, these distributions are different at the separate time points; generally, OTMs are smallest (least damage) before exposure to the chemical, largest (most damage) immediately after exposure and somewhere in-between after a longer recovery time.

We computed empirical quantiles of the OTM for each cell line at each of the three time points and then derived a single-number summary w_{ij} to tie these three quantile vectors together for cell line i and exposure j. The summary measure $w_{ij} \in (0, 1)$ is the value that minimizes

$$\sum_{h=17}^{31} |w_{ij} Q_{ij,N,h} + (1 - w_{ij}) Q_{ij,L,h} - Q_{ij,A,h}| \qquad (5)$$

Here, $Q_{ij,N,h}$ indicates the $h/32^{th}$ quantile for the i^{th} cell line's OTM distribution at the '**N**o treatment' time, with corresponding quantities for the '**L**ater' time point and the 'immediately **A**fter' time point for the j^{th} exposure. The use of only the higher quantiles reflects our desire to learn more about the extremes of DNA repair. We used a logit transform to derive our final response $y_{ij} = log(\frac{w_{ij}}{1-w_{ij}})$; this is appropriate for the assumptions of the model. Negative values of the response indicate that the OTM distribution long after treatment is closer to the distribution right after treatment; positive values indicate that the 'long after' distribution is closer to the distribution before treatment.

SNPs in genes thought to be associated with some aspect of DNA repair were genotyped, leading to data on 49 428 individual SNPs. Given the small number of cell lines and the fact that many individuals have two copies of the major allele for these SNPs, many of the SNP profiles were identical or had no individuals with two copies of the minor allele. We recoded the genotypes so that one indicated at most one copy of the major allele and two indicated two copies of the major allele. After recoding, we reduced the predictor set to those SNPs with distinct profiles, leaving 23 210 SNPs for analysis.

Table 1. Details for SNPs included in the final CTF model for the molecular epidemiology data

Gene	SNP	Position
IGFBP5	RS11575170	217256085
TGFBR3	RS17880594	92118885
CHC1L	RS9331997	47986441
XPA	RS3176745	99478631

Table 2. Comparison of MSPE, 95% coverage proportion and mean computation time for different methods applied to molecular epidemiology data

Metric	CTF	RF	QRF	BART
MSPE	0.263	0.353	–	0.425
95% Coverage	0.961	–	0.928	0.817
Time (s)	3317	80	88	2343

We used leave-one-out cross-validation to assess the performance of CTF against the three competitors RF, QRF and BART. We ran the variable selection chain for 5000 burn-in iterations and computed inclusion probabilities from 10 000 samples. We ran the MCMC chain for 40 000 burn-in iterations and retained a sample of 20 000 iterations. Autocorrelation diagnostics indicated an effective sample size of 15 000. We used the same burn-in and posterior sample sizes for BART. As in the simulation study, we used the results from a first run of RF to seed a final run of RF.

CTF showed consistent selection of the treatment (H_2O_2 or MMS) as the most important predictor and selected a set of four SNPs (IGFBP5, TGFBR3, CHC1L and XPA) as predictors; information about these SNPS is summarized in Table 1. In contrast, RF chose the treatment variable in only 56 of the 180 cross-validation scenarios and did not consistently identify any other predictors. Comparison with the competitor methods showed patterns similar to the simulation study; Table 2 compares the results from each method. The interactions between the treatment and the various SNPs may be weak enough that they do not contribute to the same elevated MSPE that RF demonstrated in the simulation study. Even though the MSPE for RF was close to that for the CTF, the CTF was able to achieve lower MSPE while not sacrificing coverage performance. This improved performance offsets the CTF's higher computational time requirement. Figure 3 shows estimated conditional densities with 95% credible intervals from the full dataset given varying levels of the treatment and of the IGFBP5 SNP while holding the other three SNPs at

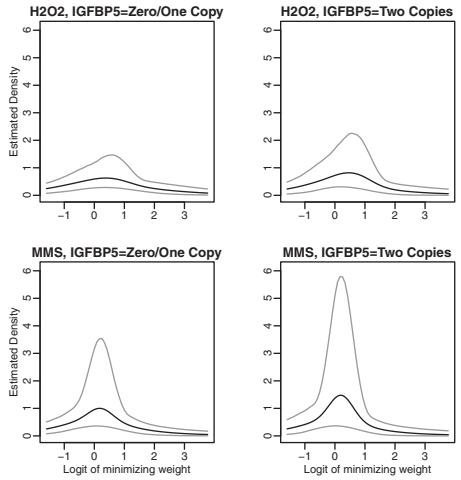

Fig. 3. Selected conditional densities given different exposures and different number of copies of the dominant allele at the IGBP5 SNP, holding all other SNPs at the 0/1 level. Heavy black lines show the mean conditional density and gray lines show the 95% credible interval

Table 3. Summary of conditional distribution characteristics

Profile	Mean	Variance	90^{th} %ILE
H2O2, IGFBP5 = 0/1	0.226	11.39	2.65
H2O2, IGFBP5 = 2	0.156	7.31	2.25
MMS, IGFBP5 = 0/1	0.023	9.76	2.07
MMS, IGFBP5 = 2	−0.023	6.11	1.88

the 'Zero/One Copy' level, and illustrates how the conditional density changes in more than the conditional mean when the predictor vector changes. In this case, the interaction between MMS treatment and two copies of the major allele for this IGFBP5 SNP tightens the density markedly, although it has a more muted impact on the conditional mean. The change is less dramatic under the exposure to H_2O_2. Here, the shift in the mean response as treatment and genetic profile change is less interesting than the difference in conditional variance; under treatment with H_2O_2, the mean response is slightly different than under treatment with MMS, but the tail probabilities are noticeably different. Table 3 summarizes these differences in conditional mean, conditional variance and conditional 90^{th} percentile for each scenario. As suggested in Figure 3, the medians of the conditional densities given the exposure (H_2O_2 or MMS) are close, but in the tail of the distribution (the 90th percentile), the separation between the estimated quantile curves is larger. This varying shift in the 90th percentile reflects the interaction between the exposure and the level of the IGFBP5 SNP.

5 CONCLUSION

We have presented a novel method for flexible conditional density regression in the common case of a continuous response and categorical predictors. The simulation study and real data example suggest that this conditional tensor factorization method can have better performance than other modeling tools when there is substantial interaction between the predictors of interest. The CTF does have a higher computational time requirement than the competitor methods, but the improvement in prediction accuracy and coverage still make the CTF an attractive method.

A particularly appealing aspect of the CTF is predictor selection, which finds low-dimensional structure in the high-dimensional predictor set. This reduction to more parsimonious models yields a succinct description of the ways in which the phenotype varies given exposure and SNPs. Finally, a distinct advantage of the CTF is its ability to produce conditional density estimates. This property of the CTF provides insight beyond a simple conditional expectation and makes it possible to answer more complex questions about the relationship between the response and the predictors.

Funding: This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences, Z01 ES049032. David Kessler's work was partially supported by National Institute of Environmental Health Sciences training grant T32ES007018. David Dunson's work was supported by Award Numbers R01ES017240 and R01ES017436 from the National Institute of Environmental Health Sciences.

Conflict of Interest: none declared.

REFERENCES

Bishop,C. and Svensén,M. (2003) Bayesian hierarchical mixtures of experts. In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. pp. 57–64.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chipman,H. *et al.* (2006) Bayesian ensemble learning. In: *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, Massachusetts, USA, pp. 265–272.

Chipman,H. *et al.* (2010) BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, **4**, 266–298.

Chu,W. and Ghahramani,Z. (2009) Probabilistic models for incomplete multi-dimensional arrays. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, Florida, USA*.

Chung,Y. and Dunson,D. (2009) Nonparametric Bayes conditional distribution modeling with variable selection. *J. Am. Stat. Assoc.*, **104**, 1646–1660.

Cowell,R. *et al.* (1999) *Probabilistic Networks and Expert Systems*. Springer, New York, USA.

De Iorio,M. *et al.* (2004) An ANOVA model for dependent random measures. *J. Am. Stat. Assoc.*, **99**, 205–215.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Society B (Methodological)*, **39**, 1–38.

Dunson,D. and Park,J. (2008) Kernel stick-breaking processes. *Biometrika*, **95**, 307–323.

Dunson,D. and Xing,C. (2009) Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Stat. Assoc. (Theory and Methods)*, **104**, 1042–1051.

George,E. and McCulloch,R. (1997) Approaches for Bayesian variable selection. *Stat. Sin.*, **7**, 339–373.

Griffin,J. and Steel,M. (2006) Order-based dependent Dirichlet processes. *J. Am. Stat. Assoc.*, **101**, 179–194.

Hannah,L. *et al.* (2011) Dirichlet process mixtures of generalized linear models. *J. Mach. Learn. Res.*, **12**, 1923–1953.

Hoff,P. (2011) Hierarchical multilinear models for multiway data. *Comput. Stat. Data Anal.*, **55**, 530–543.

Jara,A. and Hanson,T. (2011) A class of mixtures of dependent tail-free processes. *Biometrika*, **98**, 553–566.

Jordan,M. and Jacobs,R. (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.*, **6**, 181–214.

Langseth,H. *et al.* (2012) Inference in hybrid Bayesian networks with mixtures of truncated basis functions. In: *Sixth European Workshop on Probabilistic Graphical Models*, pp. 171–178.

Lauritzen,S. (1992) Propagation of probabilities, means, and variances in mixed graphical association models. *J. Am. Stat. Assoc.*, **87**, 1098–1108.

Meinshausen,N. (2006) Quantile regression forests. *J. Mach. Learn. Res.*, **7**, 983–999.

Moral,S. *et al.* (2001) Mixtures of truncated exponentials in hybrid Bayesian networks. In: *Proceedings of the Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning Under Uncertainty (ECSQARU 2001)*. New York, USA, pp. 156–167.

Muller,P. *et al.* (1996) Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79.

Olive,P. *et al.* (1991) DNA double-strand breaks measured in individual cells subjected to gel electrophoresis. *Cancer Res.*, **51**, 4671–4676.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, California, USA.

Reich,B. and Fuentes,M. (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Stat.*, **1**, 249–264.

Reich,B. *et al.* (2011) Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics*, **67**, 886–895.

Rodriguez,A. *et al.* (2009) Bayesian hierarchically weighted finite mixture models for samples of distributions. *Biostatistics*, **10**, 155–171.

Shahbaba,B. and Neal,R. (2009) Nonlinear models using Dirichlet process mixtures. *J. Mach. Learn. Res.*, **10**, 1829–1850.

Tokdar,S. *et al.* (2010) Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.*, **5**, 319–344.

Tucker,L. (1966) Some mathematical notes on 3-mode factor analysis. *Psychometrika*, **31**, 279.

Waterhouse,S. *et al.* (1996) Bayesian methods for mixtures of experts. In: *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, Massachusetts, USA, pp. 351–357.

Yang,Y. and Dunson,D.B. (2012) Bayesian Conditional Tensor Factorizations for High-Dimensional Classification. *Working Paper*. Duke University, Durham, USA.

Xu,Z. *et al.* (2012) Infinite Tucker decomposition: nonparametric Bayesian models for multiway data analysis. In: *Proceedings of the 29th International Conference on Machine Learning, Princeton, New Jersey, USA*.