Quantitative Marketing and Economics (2019) 17: 415–437
https://doi.org/10.1007/s11129-019-09216-4

VANISH regularization for generalized linear models

Oliver J. Rutz[1] · Garrett P. Sonnier[2]

Received: 22 July 2018 / Accepted: 7 July 2019 / Published online: 14 August 2019

Abstract
Marketers increasingly face modeling situations where the number of independent variables is large and possibly approaching or exceeding the number of observations. In this setting, covariate selection and model estimation present significant challenges to usual methods of inference. These challenges are exacerbated when covariate interactions are of interest. Most extant regularization methods make no distinction between main and interaction terms in estimation. The linear VANISH model is an exception to these methods. The linear VANISH model is a regularization method for models with interaction terms that ensures proper model hierarchy by enforcing the heredity principle. We derive the generalized VANISH model for nonlinear responses, including duration, discrete choice, and count models widely used in marketing applications. In addition, we propose a VANISH model that allows to account for unobserved consumer heterogeneity via a mixture approach. In three empirical applications we demonstrate that our proposed model outperforms main effects models as well as other methods that include interaction terms.

Keywords Non-linear marketing models · High-dimensional data, · Interactions, regularization methods · Bayesian methods

1 Introduction

As data capture and storage costs fall, marketers are increasingly able to collect and utilize data on a multitude of consumer-firm contacts and customer activities (i.e., email, web visits, call center contacts, posts, tweets, purchases, etc.). The result of this explosion in information, much of it unstructured, has resulted in data that are aptly

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11129-019-09216-4) contains supplementary material, which is available to authorized users.

✉ Oliver J. Rutz
orutz@uw.edu

[1] University of Washington, Box 353226, Seattle, WA 98195, USA

[2] The University of Texas at Austin, 1 University Station, Austin, TX 78712, USA

characterized as high dimensional. In most applications, the computational burden imposed by large numbers of observations can be addressed by sampling and/or parallel computing (Bumbaca et al. 2017). A more problematic situation occurs when the number of predictors becomes large relative to the number of observations. This is the so-called "large p, small n" problem, where the number of predictors approaches or exceeds the number of observations. Allowing for interaction effects in this setting aggravates the problem considerably. However, such effects may likely be useful in improving prediction in typical marketing settings. For example, as textual covariates convey meaning, more information may be contained in the interactions of these covariates. As including interaction effects further exacerbates the "large p, small n" problem some type of regularization of the likelihood or dimension reduction method will be required for model estimation.[1]

In the context of a "large p, small n" problem with interactions a natural issue to consider is how any regularization treats the main effects and interaction terms. Interaction terms included in a linear model without the corresponding main effects can lead to issues in model estimation (Peixoto 1990; Nelder 1998). For example, if main effects are excluded from a model with interactions, the model is not invariant under a coding transformation, such as mean centering the independent variables in the model. Regularization approaches for linear models, such as Ridge Regression and the LASSO model, are well known. In principle these models can handle interaction terms but penalize them equally and are equally likely to admit an interaction term or a main effect. Such approaches may violate the principle of a well-formulated polynomial model (Peixoto 1990). Recently, the VANISH model (Radchenko and James 2010) addresses the issue of treatment of main and interactions terms in regularized linear regression models by developing a model that follows the heredity principle (Nelder 1998). The degree of penalization on the interaction terms depends on whether the main effects are already present in the model.[2] While VANISH regularization is well understood for linear models many marketing applications involve the study of non-linear phenomena. In addition, for many marketing applications accounting for unobserved heterogeneity is important.

The main contribution of this paper is twofold. First, we extend the VANISH regularization approach to accommodate generalized linear response models (GLM) of interest to marketing academics and practitioners, including hazard, discrete choice, and count models. Second, we show how to accommodate unobserved heterogeneity in a VANISH regularization for generalized linear models. We adopt a Bayesian approach to inference which readily adapts to nonlinear marketing response models and also accommodates the tuning parameters in the model hierarchy permitting simultaneous estimation of all model parameters. In a frequentist setting the tuning parameters of any regularization approach need to be inferred using cross-validation techniques subsequent to parameter estimation. We demonstrate the superior predictive performance of our proposed VANISH regularized GLM with three empirical marketing datasets.

[1] For example, say a model has 100 parameters. When adding only first-level interactions, one needs to estimate 100 main effects and 4,950 interaction effects. Even if enough observations are available to estimate the main effects, adding the interaction effects will almost certainly result in a "large p, small n" problem.

[2] VANISH refers to Variable Selection using Adaptive Nonlinear Interaction Structures in High Dimensions (Radchenko and James 2010).

Our first empirical application extends the linear VANISH model to the problem of modeling the time until a customer life event in a customer relationship setting. A life-event is a customer behavior that is not directly linked to a customer's interaction with the firm but changes the products/services of interest to a customer in a structural way. For example, it is well known that the birth of the first child provides insurance companies an excellent opportunity to sign up consumers for life-insurance. Using a novel customer-level and predictor-rich dataset we show how to improve prediction of a life event with our proposed model. Our data provided by the Wharton Customer Analytics Initiative (WCAI) contain 101 monthly metrics describing the customer-firm interaction over a span of 12 months. We use these metrics together with first-order interactions to predict whether a life-event has occurred. We find that our proposed approach outperforms standard methods of forecasting life-events based on main effects only as well as other regularization approaches that allow for interactions.

In a second application we extend the linear VANISH model to a discrete choice panel data setting and show how to account for unobserved heterogeneity. We apply our modeling approach to a data set from Twitter that contains repeated user level observations (i.e., panel data) on retweeting behavior by non-professional (i.e., not corporate or managed accounts) Twitter users. We consider the influence of message content on retweeting by using a text mining approach to quantitatively represent message content. Our VANISH approach accommodates both interactions in the textual covariates and unobserved user heterogeneity. We show first that our proposed heterogeneous VANISH choice model improves in-sample and out-of-sample model performance compared to homogenous and heterogeneous models that consider only main effects. Our VANISH choice model also out-performs other regularization approaches that consider interactions. In addition we find that accounting for unobserved heterogeneity in the VANISH choice model improves in- and out-of-sample performance over a homogenous VANISH choice model.

In the third empirical application we consider consumer response to paid search text ads (i.e., the quantity of clicks on the ad) for a mobile app. The ads are served in response to a web search. We model the count of ad clicks with a Poisson response model. Response to the ad is modeled as a function of main and interactions effects of the coded text elements of both the ad and the search terms as well as the position of the ad on the search engine results page and the cost-per-click (CPC) (which serves as a proxy for keyword popularity). We find that our VANISH regularization approach results in superior in-sample and out-of-sample fit relative to main effects only models as well as other regularization approaches that consider interactions. We use our results to determine the best fit between keywords and ads for an existing campaign and demonstrate how the campaign would improve if the keyword-text ad match would be optimized based on our results.

The remainder of the paper is structured as follows. We begin with a brief overview of regularization methods and introduce our VANISH regularization approach to generalized linear models. We introduce our first empirical application, the data on customer life-events, and discuss how the generalized VANISH model applies to modeling event duration in customer-relationship management settings. We follow with a discussion of the results. We introduce our second empirical application, the retweet data, and discuss how the VANISH regularization model applies to discrete choice with unobserved heterogeneity. We follow with a discussion of the results.

Lastly, we introduce our third empirical example, paid search click response. We discuss how incorporating the unstructured text data necessitates our VANISH regularized Poisson model. We then discuss the results. In the penultimate section we discuss the implications of our framework for marketing in predictor-rich and unstructured data settings. We conclude by noting some limitations of our approach and discussing future research.

2 VANISH regularization for linear and generalized linear models

We begin with a brief overview of extant approaches to the "large p, small n" problem. In many, if not most, empirical problems the number of observations available significantly exceeds the number of model parameters. In this setting, the asymptotic properties of estimators are well understood. In a linear regression model, for example, the ordinary least squares estimates of the model coefficients converge in distribution to their true values as the sample size increases. However, when the number of predictors approaches or exceeds the number of observations the OLS estimator is infeasible. In a survey of approaches to the "large p, small n" problem, Naik et al. (2008) broadly identify methods to address estimation problems in this setting such as inverse regression methods, factor analytic approaches and regularization methods. These methods are essentially differentiated by whether or not they solve the problem by reducing the dimensionality of the predictor matrix or regularizing the likelihood function via some penalty term.

Dimension reduction methods sacrifice some of the information in the full dimensional space in exchange for a lower dimensional space. Furthermore, it is necessary to account for the measurement error in the projection of the higher dimensional space to the lower dimensional space. Regularization methods, on the other hand, do not require any dimension reduction methods and ably use all of the information in the data. Estimation proceeds by imposing a penalty on the likelihood to account for the fact that the matrix of predictors becomes increasingly ill conditioned as the number of predictors increases relative to the sample size. For linear problems, the regularization approach is generally given by:

$$\begin{aligned} y &= \mu + X\beta + \varepsilon \quad \varepsilon \sim N\left(0, \sigma^2\right) \\ \beta_P &= \arg\min_{\beta} \left(y - X\beta\right)' \left(y - X\beta\right) + P, \end{aligned} \tag{1}$$

where y is the response variable, $y = y - \bar{y}, X$ is a matrix of predictors, β and σ^2 are parameters to be estimated and P is the penalty function.

Choice of the penalty function yields different versions of the regularized regression model. Popular penalty functions are Ridge (Tikhonov and Arsenin 1977) where $P = \lambda \sum_{i=1}^{P} \beta_i^2$, LASSO (Tibshirani 1996) where $P = \lambda \sum_{i=1}^{P} |\beta_i|$, and Elastic Net (EN) (Zou and Hastie 2005) where $P = \lambda_1 \sum_{i=1}^{P} |\beta_i| + \lambda_2 \sum_{i=1}^{P} \beta_i^2$.[3] In principle, any of these models can

[3] LASSO refers to Least Absolute Shrinkage and Selection Operator.

include interaction terms in the covariate matrix, X. However, none of these models allows distinguishing between main and interaction effects, treating lower order and higher order terms equally without regard to whether the resulting polynomial is well formed (Peixoto 1990) or follows the heredity principle (Nelder 1998). The VANISH model belongs to the class of regularization approaches but differs in that it prioritizes proper model hierarchy by enforcing the heredity principle (Nelder 1998). The degree of penalization on the interaction terms depends on whether the main effects are already present in the model. Rutz et al. (2017) show that the VANISH regularization for linear models results in superior predictive performance compared to LASSO and Elastic Net.

Models that relate censored, categorical, ordinal or count responses to covariates are commonplace in the academic marketing literature. Classic examples include hazard models of inter-purchase durations or service relationship durations, discrete choice models for category incidence and brand choice, and count models of purchase quantities. As marketers increasingly make use of unstructured data, feature engineering techniques can produce data sets with large numbers of covariates. Apart from the issue of unstructured data, consumers can now engage with a company across a variety of channels including a physical store, a company website, a mobile application, a Twitter account, a physical catalog, or email. Data on multiple touchpoints can also lead to large numbers of covariates. If interaction terms are of interest the "large p, small n" problem can be significantly exacerbated. We show how VANISH regularization can be extended to nonlinear response models for high dimensional marketing data with interactions.

To specify VANISH regularization for a GLM assume that the systematic portion of the model takes the following form:

$$\theta = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j<k} x_{ij}x_{ik}\beta_{jk} \; for \; i = 1, ..., n, j = 1, ..., k. \tag{2}$$

The model is completed by a probability distribution for the observed data y (typically induced by a distribution on a random error term,ε, added to the model) and a function $f(\theta)$ that specifies the link between the systematic and random components of the model. Different probability and link functions lead to different models. For example, a normal distribution for $y$coupled with an identity link function leads to the linear VANISH model. A binomial distribution for y_i coupled with a logit, probit, or complementary log-log link function yields different binary choice VANISH models. A Poisson distribution for y coupled with a log link function yields a Poisson VANISH. If the count data are units of time till an event and the data are censored a hazard VANISH model can be specified.

To implement the generalized VANISH model described in part by Eq. (2), $\frac{1}{2}(p^2 + p)$ beta parameters need to be estimated. Assume, for example, 500 observations can be described by a modest numbers of covariates, say 40. The number of parameters to be estimated in this example is 820. Thus, estimating the proposed model with standard methods of inference is problematic. Furthermore, given our desire to adhere to the hereditary principle we require the VANISH regularization approach. In our Bayesian framework an informative prior for β is needed to implement this approach.

We will now discuss the penalty function of the VANISH model and show how the model allows for different treatment of main and interactions effects.

Standard regularization approaches (e.g., Ridge Regression or LASSO) treat main and interactions effects equivalently. This gives no adherence to the hereditary principle (Nelder 1998) and readily allows interaction terms regardless of the status of the corresponding main effects. The VANISH penalty automatically adjusts the degree of shrinkage on the interactions depending on whether the main effects are already present in the model. An added benefit of the VANISH penalty is the ease with which interaction terms can enter if the corresponding main effects have already been added. The VANISH penalty is given by:

$$P_{VANISH} = \lambda_1 \sum_{j=1}^{p} \left(\beta_j^2 + \sum_{k:k\neq j}^{p} \beta_{jk}^2 \right)^{1/2} + \lambda_2 \sum_{j=1}^{p} \left(\sum_{k=j+1}^{p} |\beta_{jk}| \right), \tag{3}$$

where λ_1 reflects the weight of the penalty for each additional predictor included in the model, λ_2 reflects the additional penalty on the interaction terms, p is the number of predictors, and $\beta$are parameter vectors to be estimated. Note that while in the LASSO the prior can be expressed as a normal distribution, the VANISH prior cannot be expressed as a normal distribution. Rutz et al. (2017) show how to derive the VANISH penalty via a Laplace transform. The VANISH prior is expressed as

$$\beta \propto \exp\left[-\frac{\lambda_1}{\sigma} \sum_{j=1}^{p} \left| \left(\sum_{k=j+1}^{p} |\beta_{jk}| \right) \right| - \frac{\lambda_2}{\sigma} \sum_{j=1}^{p} \left| \left(\beta_j^2 + \sum_{k:k\neq j}^{p} \beta_{jk}^2 \right)^{1/2} \right| \right]. \tag{4}$$

A benefit of the Bayesian approach to estimation is the ability to estimate the penalty parameters λ_1 and λ_2 directly in the sampler whereas in the classical framework they are estimated using cross-validation. To complete the model we specify gamma priors on the penalty parameters λ_1 and λ_2:

$$\lambda_1^2 \sim gamma(c, d) \text{ and } \lambda_2^2 \sim gamma(c, d), \tag{5}$$

where c and d are parameters chosen to ensure an uninformative prior. We now turn attention to our three empirical applications, beginning with the multi-touch data on customer life changes and financial activity.[4]

3 Predicting life changes from financial activity

3.1 Customer relationship management and life events

Managing the customer relationship across the phases of acquisition, development and retention has long been of interest to marketing academics and practitioners. Increasingly, rich and large data sets are available to build models and inform marketing

[4] Details on the sampler can be found in the Appendix. More details on the derivations of the full conditional distributions of the VANISH parameters can be found in the Web Appendix.

strategies in terms of all three phases. Conceptually, one can create response models estimated on customer-specific *characteristics* (i.e., age or income) or customer *interactions* with the firm (i.e., how a customer executes financial transactions with the firm). Additionally, marketers may attempt to leverage customer behavior that is not directly linked to a customer's interaction with firm. For example, the birth of a first child can provide insurance companies an opportunity to offer consumers life-insurance products. Such an event is called a life-event. Our application considers a financial service company interested in predicting a relevant life-event based on its interaction data with the customer.

3.2 Data

The data are provided by the Wharton Customer Analytics Initiative (WCAI). The firm is a financial service provider interested in forecasting a specific event in the lives of their customers. Due to confidentiality we cannot specify the firm or the life-event in question. The life-event in question signals the end of a period in the customer's life (e.g., marriage, terminating employment, etc.). It is important to note that the life-event in question does not terminate the customer's tenure with the firm. As with the life-events described above (e.g. pregnancy) the firm has no influence on whether and when the specific life-event in question occurs. However, the focal firm in this study does get to observe and record the event at some point after it occurs. From the perspective of the firm, the life-event in question dramatically expands the type of services the customer can buy from the firm. Knowledge of the timing of such a life event would give the firm a powerful tool to target its existing customer base at the right time, potentially before a customer starts to search for these products/services in the marketplace by approaching competitors.

The data consists of observations on 98,088 firm customers over January 2012 to January 2013. Of these, 17,546 customers have an observed life-event in the time span. For each customer, we have 12 monthly observations that describe customer interactions with the firm. These interactions are recorded as three distinct classes of information. First, for each customer we have information on the products and services used in each month. The firm tracks 27 different metrics representing its product and service offerings. These data are called "products" going forward. Second, for each month the firm tracks its contact with the customer. The metrics include customer-initiated inbound contacts as well as firm-initiated outbound contacts. The firm tracks 28 different contact metrics that we will call "contacts" going forward. Lastly, the firm tracks so called business process handles (BPH). These represent whether a certain business process is ongoing. Such processes could be providing the customer with a quote or changing the customer's product and service portfolio. The firm tracks 46 different monthly BPH metrics. In sum, for each customer for each month we have 101 different metrics describing the customer-firm interaction.

3.3 Model-free evidence

Before proceeding to the model we present some model free evidence aggregating over products, contacts and BPH resulting in three monthly metrics per customer. Using the 17,546 customers with a valid recorded life-event we calculate the percent change for a 1, 2 and 3 month window before the life-event for the three metrics. We find that for

BPH and contacts there is a large increase in activity compared to the average of the activity before the start of the window. We do not find the same for products. For BPH, we find that there is about a 200% increase in the last month before the life-event (136% increase for the 2 month window and 123% increase for the 3 month window). For contacts, we find that there is a 126% increase in the last month before the life-event (91% increase for the 2 month window and 89% increase for the 3 month window). We find no difference for products. In sum, it appears the customer's interactions with the firm are changing significantly before the life-event.

3.4 A model to forecast life-events

Given that we are interested in the duration until the life event, that over our observation period some of the units do not experience the event, and that we have a number of predictors that may affect the waiting time, hazard models well suited to our task. Let the random variable t denote the time till the life event. We model the propensity for a life-event, $h_i(t, x_{it})$, using a proportional hazard approach

$$h_i(t, x_{it}) = h_0(t, \alpha, \gamma)\psi(\theta_{it}) \tag{6}$$

where i is customer, $h_0(t, \alpha, \gamma)$ is the baseline hazard and $\psi(\theta_{it})$ is the customer-specific multiplicative variation of the baseline hazard. We use a Weibull baseline hazard which defines $h_0(t, \alpha, \gamma)$ as follows

$$h_0(t, \alpha, \gamma) = \alpha\gamma(\gamma t)^{\alpha-1} \tag{7}$$

where α and γ are parameters to be estimated. As is well known, the model is flexible in the sense that if $\alpha = 1$ the model reduces to the exponential hazard with constant risk over time. Values of α greater or less than 1 correspond to increasing or decreasing risk over time, respectively.

The proportional hazard model allows for time varying covariates to influence the survival time. We model main effects and first order interactions as follows

$$\begin{aligned} \psi(\delta_{it}) &= \exp(\theta_{it}) \\ \theta_{it} &= \sum_{j=1}^{p} x_{ijt}\beta_j + \sum_{j<k} x_{ijt}x_{ikt}\beta_{jk} \end{aligned} \tag{8}$$

where x_{it} are available metrics that capture the firm-customer relationship, and β is a set of parameters to be estimated. Rather than a normal prior we use the VANISH prior for β as described in (4).

3.5 Results

To illustrate the power of the VANISH approach we estimate our model on an intentionally small sub-sample of 300 randomly chosen customers for the time period February 2012 till January 2013. As we are interested in prediction we use the values of the product, contact, and BPH metrics in the previous month. We exclude the first month of the data, as we do not have predictors available in the data. We include all

observed first-order interactions for the 101 metrics, resulting in the need to estimate a total of 1,903 predictors describing the effect of the customer-firm interactions. Our goal here is prediction of the life event, not causal inference. Table 1 reports in-sample fit as measured by the Deviance Information Criterion (Spiegelhalter et al. 2002). We compare our VANISH Hazard model with interactions to a Hazard model using main effects alone as well as a hazard model that includes interactions via a LASSO prior.[5] The LASSO prior does not differentially penalize main effects and interaction terms. We find the VANISH Hazard model provides the best fit to the data in-sample.

The goal of the firm is to predict the life-event occurring based on the customer-firm interaction data alone. We show that our VANISH Hazard model is well suited to this task. We start by choosing 10,000 customers at random for the hold-out task. Naturally, these do not include the 300 customer used for estimation. For the 10,000 hold-out customers we forecast the life-event and compare our forecast with their actual behavior. We average over 100 forecasts for each customer. The results appear in Table 2. We find that the VANISH Hazard approach fits the data better than a main effects only hazard model and a LASSO Hazard model in terms of hold-out mean squared error (MSE) and holdout mean absolute error (MAE). For the sake of comparison we also estimate the out-of-sample performance for a hazard multiple adaptive regression tree (MART). The MART ably handles large numbers of parameters since it performs optimization efficiently in gradient space using a greedy algorithm (Yoganarasimhan 2018). The hierarchical structure of MART naturally models interactions as the response to one predictor variable depends on predictors higher in the tree.[6] Our proposed generalized VANISH yields better out-of-sample performance compared with the Hazard MART as well.

Lastly, we investigate the performance of our models for Type I and Type II errors by splitting the holdout data into observations where the life-event has not occurred and observations where the life-event has occurred. We then examine the predictive performance of the model in these two settings. The results appear in Table 3. In terms of correctly forecasting an observed life-event, the VANISH Hazard outperforms the benchmark Hazard model as well as the LASSO Hazard and MART Hazard models with substantial improvement in forecast MAD and MSE. From the firm's perspective, identifying customers in time before the life-event occurs is critical to offer the products and service changes that will accompany the life-event. The ability to forecast the life-event with greater precision enables the firm to move early in an effort to acquire new business from existing customers with the life-event. However, erroneously targeting customers less likely to respond because they have not experienced the life-event is costly. In this setting, our proposed model also outperforms the basic benchmark Hazard model as well as the LASSO and MART Hazard models in terms of forecast

[5] The LASSO prior is $\pi(\beta|\sigma) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} \exp\left(-\frac{\lambda|\beta_j|}{\sigma}\right)$. Note that the LASSO model only requires one tuning parameter, λ, and one set of latent parameters, τ. Estimation proceeds similarly to estimation using a VANISH prior as detailed in the Appendix. Also see Park and Casella (2008) for a full Bayesian treatment of the linear LASSO.

[6] We estimate the MART version of each of our models (hazard, choice and count) using the gbm package in R. The gbm algorithm is a boosting algorithm which creates a sequence of simple trees where each successive tree is built for predicting the residuals of the preceding tree. Thus, at each step of the algorithm a simple partitioning of the data is determined and the deviations of the observed values from the respective means (residuals for each partition) are computed. The next tree will then be fitted to those residuals to find another partition that will further reduce the residual (error) variance given the preceding sequence of trees.

Table 1 In-sample fit

Model	DIC[a]
Hazard	−313.02
LASSO Hazard	−269.93
VANISH Hazard	−257.55

[a] Deviance information criterion based on Spiegelhalter et al. (2002)

MAE and MSE among observations where no life-event is observed. From the firm's perspective, this is significant as this means that the number of customers erroneously targeted decreases significantly.

4 Predicting rebroadcasting behavior on twitter

4.1 The relevance of rebroadcasting

On social platforms such as Facebook and Twitter recipients of a message or content may "repost", "share", or "retweet" the content. A growing number of scholars have examined the dynamics of content rebroadcasting. Most research emphasizes the use of network-centric features in modeling rather than using the message content itself (e.g., Cheng et al. 2014; Petrovič et al. 2011; Zaman et al. 2014; Bakshy et al. 2011). Few studies have evaluated message content (e.g., Hong et al. 2011; Kleinberg 2014) and those that have conclude the predictive value of content is low (Kleinberg 2014). However, understanding the predictive power of message content is especially important. Agents have little to no control over the structure of social networks in the audiences that they seek to reach. They do, however, have control over the content of their messages. In this application we use message content to predict message rebroadcasting. We use the **L**inguistic **I**nquiry and **W**ord **C**ount (LIWC) (Pennebaker 2011) program to quantitatively represent the message content received by a user. To account for interactions between the content metrics we use a VANISH regularization approach. As we observe multiple rebroadcast opportunities per user we incorporate unobserved

Table 2 Out-of-sample fit

Model	MSE[a]	MAE[b]
Hazard	11.32	1.64
LASSO hazard	10.73	1.52
Hazard MART[c]	9.77	1.35
VANISH hazard	6.72	1.02

[a] Mean squared error (using 100 forecasts, 10,000 Holdout IDs)

[b] Mean absolute error (using 100 forecasts, 10,000 Holdout IDs)

[c] 10,000 trees, 5 fold cross-validation

Table 3 Out-of-sample forecast comparison

Model	MSE[a]		MAE[b]	
	Life event occurs	No life event occurs	Life event occurs	No life event occurs
Hazard	7.47	31.60	1.01	4.94
LASSO hazard	7.10	29.68	0.93	4.64
Hazard MART[c]	6.70	25.92	0.79	4.25
VANISH hazard	3.19	25.26	0.43	4.12

[a] Mean squared error (using 100 forecasts, 10,000 Holdout IDs)

[b] Mean absolute error (using 100 forecasts, 10,000 Holdout IDs)

[c] 10,000 trees, 5 fold cross-validation

user heterogeneity into the VANISH regularization approach via a mixture model. We compare and contrast model performance across models that vary in terms of using interaction terms and accounting for unobserved heterogeneity.

4.2 Data and model

We utilize a data set on the broadcasting and rebroadcasting behaviors of a set of Twitter users over a four-week period. The users are randomly selected from a list of users present in Twitter's *spritzer* dataset, which is a 1% random sample of Twitter activity.[7] The data include users with no more than 3,000 followers or friends, who tweeted solely in English, and tweeted at least 10 times in the previous 2 weeks. Additionally, corporate accounts and celebrity fan accounts (i.e., professional and pseudo-professional Twitter accounts devoted exclusively to one purpose) are excluded. We collect all of the tweets and retweets from a set of users, as well as all of the tweets from every friend of each of these users, over 30 days. In our final analyses, we exclude all users who had not tweeted *and* retweeted during the observation period, and randomly sample 100 users from the remaining subset of users. For each user, we randomly sample 20 tweets and record the retweet status of each user-tweet pair.

We also observe the content of the messages being broadcasted and rebroadcasted. To process the data, we use the LIWC software (Pennebaker 2011) to create a vector of content features for each tweet. LIWC builds values for each of its features based on the presence and prevalence of certain pre-specified terms. Note that there are many other ways to mine text content. We are not aiming to propose a "best" way to mine content or provide an exhaustive catalogue of methods. Rather, our goal is to create a dataset to illustrate our VANISH regularization approach to account for potential interactions between the content features. Each data point describes the retweet decision of an individual user. For each of our

[7] We thank Maytal Saar-Tsechansky of UT Austin for providing access to the data and Samuel Blazek for providing research assistance in processing the data.

100 users we have 20 retweet decisions in total. We use 18 for model estimation (1,800 datapoints) and 2 for holdout (200 datapoints). We have 68 mean effects and 2,278 interaction effects in terms of the LIWC content features.

For our sample of $i = 1,\ldots,n$ users we observe a set of $l = 1,\ldots,k_i$ retweet opportunities per user i. Let y_{il}^* denote latent utility for a retweet of user i at opportunity l.

$$\begin{aligned} y_{il}^* &= \theta_{il} + \varepsilon_{il}, i = 1, ..., n \quad \varepsilon_{il} \sim EV(0,1) \\ \theta_{il} &= \sum_{j=1}^{p} x_{ilj}\beta_j + \sum_{j \neq k} x_{ilj}x_{ilk}\beta_{jk} \end{aligned} \tag{9}$$

where x_{ilj} is the LIWC content feature covariate j for user i at opportunity l, p is the number of LIWC features, and β is a vector of parameters to be estimated.

Given that we observe multiple retweet opportunities per user it is natural to consider accounting for unobserved user heterogeneity.[8] While the VANISH prior is well suited to "large p, small n" estimation problems and explicitly handles interaction terms it is not clear that it is amenable to a continuous heterogeneity approach. To model unobserved heterogeneity we leverage a finite mixture approach (e.g., Allenby et al. 1998). We model the vector of response parameters β as well as the vector of VANISH prior parameters λ as arising from a finite mixture of Sdiscrete components. We can conceptualize this approach as assuming that each individual observation arises from an unknown component of the mixture,z_i, where z_1, ..., z_n are realizations of the independent and identically distributed random variables Z_1, ..., Z_n with a probability mass function

$$\Pr(Z_i = s|\pi) = \pi_s, (i = 1, ..., n \text{ and } s = 1, ..., S). \tag{10}$$

The parameters $\pi = (\pi_1, ..., \pi_s)$ are the mixture proportions which are constrained to be non-negative and to sum to 1.

The likelihood of the observed data is a finite mixture of the likelihoods of each component

$$f(y|x;\beta,\lambda,\pi) = \sum_{s=1}^{S} \pi_s f_s(y|x;\beta^s,\lambda^s) = \sum_{s=1}^{S} \pi_s f_s(y|x;\beta^s) g_s(\beta^s|\lambda^s) h_s(\lambda^s). \tag{11}$$

For the s^{th} component we model f with a binary logit likelihood conditional on segment membership

[8] Gilbride et al. (2006) suggest a heterogeneous variable selection approach applied to conjoint data that shrinks individual-level response coefficients either towards zero with a very small prior variance or a normal prior distribution dependent upon individual-level discrete variable selection parameters. Their model performs variable selection at the attribute level rather than the partworth level (i.e., the brand attribute is either selected or not versus each level of the brand attribute). While this approach allows the researcher to heterogeneously model attribute attendance at the individual level it is not, per se, an approach well suited for "large p, small n" problems. Their approach does not include a way to ensure that the number of selected parameters (i.e., the parameters not shrunk towards zero) does not exceed the number of observations. Their approach is also mute on whether and how to handle interaction terms.

$$\Pr[y_{il} = 1|Z_i = s] = \frac{\exp\left(\sum_{j=1}^{p} x_{ilj}\beta_j^s + \sum_{j\neq k} x_{ilj}x_{ilk}\beta_{jk}^s\right)}{1 + \exp\left(\sum_{j=1}^{p} x_{ilj}\beta_j^s + \sum_{j\neq k} x_{ilj}x_{ilk}\beta_{jk}^s\right)}. \quad (12)$$

The function g is the segment specific VANISH prior for β given by Eq. (4) while the function h is the segment specific gamma priors for the VANISH penalty parameters given by Eq. (5). The mixture components are modeled as arising from a Dirichlet distribution

$$\pi \sim Dir(\rho), \quad (15)$$

where ρ is a S-vector of prior hyperparameters $\rho = (\rho_1, ..., \rho_S)$.

A well-known problem with mixture models is label switching. While a unique labeling is required for inference about mixture components predictive densities are identical for all label permutations and the average over conditional predictive densities computed at each draw is invariant to label switching.[9] As we are chiefly concerned with prediction we could safely ignore the label switching problem. However, other researchers may want to conduct inference on the mixture components. Thus, we demonstrate how to use the relabeling algorithm proposed by Stephens (2000) which provides a simple and robust method to address label switching. The additional steps required to implement the relabeling algorithm are included in the Appendix.

4.3 Results

We compare our choice models of rebroadcasting behavior across the dimensions of accounting for unobserved heterogeneity and inclusion of textual interaction terms. Table 4 reports the in-sample fit statistics as measured by DIC. We estimate a homogeneous choice model, latent class choice models with 2 and 3 segments, and a choice model with continuous heterogeneity. None of these models include interaction terms due to the "large p, small n" issue that would arise. We compare these models to homogenous and latent class LASSO and VANISH choice models with 2 and 3 segments, respectively. Each of the three VANISH models outperforms its comparison model (i.e., the homogeneous, 1 segment, and 2 segment, respectively). Additionally, the 2 and 3 segment VANISH choice models out-perform the model with continuous heterogeneity but without interactions. Table 5 reports the out-of-sample hit rate. To compare out-of-sample results we also estimate a MART choice model. Similar to the in-sample performance, we find each of the three VANISH models outperforms its comparison model out-of-sample. Each of the three VANISH models also outperforms its comparison LASSO model. In terms of out-of-sample performance we find that a two segment mixture VANISH predicts the best, including when compared to a model with continuous heterogeneity and the MART choice model.

[9] We thank the editor for bringing this point to our attention.

Table 4 In-sample model fit statistics

Model	Interactions	Heterogeneity	Segments	DIC
Binary choice model	No	N/A	1	−1,146.1
Binary choice model	No	Discrete	2	−936.1
Binary choice model	No	Discrete	3	−856.4
Binary choice model	No	Continuous	N/A	−341.8
LASSO binary choice model	Yes	N/A	1	−1089.1
LASSO binary choice model	Yes	Discrete	2	−541.4
LASSO binary choice model	Yes	Discrete	3	−397.4
VANISH binary choice model	Yes	N/A	1	−816.9
VANISH binary choice model	Yes	Discrete	2	−330.7
VANISH binary choice model	Yes	Discrete	3	−185.9

5 Predicting click response in paid search advertising

5.1 Paid search

Paid search advertising is perhaps the best performing advertising vehicle of the internet economy. Extant research shows that the position of the paid search ad is a key driver of the customer's click decision (e.g., Ghose and Yang 2009; Agarwal et al. 2011; Ghose et al. 2014). While concerns over position endogeneity typically loom due to the auction such concerns can be safely ignored if the researcher is purely concerned with holdout prediction and not inference (Ebbes et al. 2011). Extant models treat keywords as a source of heterogeneity and address these differences in response across keywords by either grouping keywords into classes or estimating keyword-level parameters (e.g., Ghose and Yang 2009; Rutz et al. 2012). Rutz et al. (2017) use lab

Table 5 Out-of-sample model fit statistics

Model	Interactions	Heterogeneity	Segments	Hitrate
Binary choice model	No	N/A	1	0.55
Binary choice model	No	Discrete	2	0.67
Binary choice model	No	Discrete	3	0.64
Binary choice model	No	Continuous	N/A	0.63
LASSO binary choice model	Yes	N/A	1	0.57
LASSO binary choice model	Yes	Discrete	2	0.63
LASSO binary choice model	Yes	Discrete	3	0.65
MART choice model[a]	Yes	N/A	1	0.58
VANISH binary choice model	Yes	Discrete	1	0.60
VANISH binary choice model	Yes	Discrete	2	0.71
VANISH binary choice model	Yes	Discrete	3	0.69

[a] 10,000 trees, 5 fold cross-validation, one Segment only

data to investigate the role of the text ad in consumers' click decision. Here, we use secondary data to examine the role of textual content. We use the bag of words method to capture the counts of the words used in the copy (e.g., Salton and McGill 1983). Previous research has considered interactions in an ad-hoc manner by including indicator variables that measure whether a keyword appears in the ad in general or whether the keyword appears in the headline of the ad or the body of the ad. Our proposed approach is more general as it looks at all possible interactions between keyword predictors and ad predictors.

5.2 Data and models

Our data are from an advertiser selling a mobile app product. The campaign is comprised of 58 keywords and 28 versions of the text ad. All keywords and ads are used to advertise a single mobile app product. The 28 text ads are essentially slight textual variations of the same ad. After a search for any of the 58 keywords consumers are served one of the 28 text ads. Clicking on the ad brings the customers to a common landing page for the same mobile app product. Our data comprise of 3,371 daily observations. On average, the campaign creates 71 impressions and 5.6 clicks per day. Average cost-per-click (CPC) is $0.20 and the average position is 2.01. Bag-of-words coding (after stemming and removing stop words) results in 43 unique words representing the text in the 58 keywords and the 28 ads. Keywords, on average, contain 3.1 words and text ads, on average, contain 12.3 words.

Let the daily number of clicks be denoted by y_t. We model the daily number of clicks as a Poisson process

$$\Pr[y_t] = \frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!}. \tag{13}$$

We express θ_t as follows

$$\theta_t = \ln(\lambda_t) = x_t^{sea}\beta^{sea} + x_t^{imp}\beta^{imp} + x_t^{pos}\beta^{pos} + x_t^{cpc}\beta^{cpc} + \sum_j x_{jt}^{txt}\beta^{txt} + \sum_{j<k} x_{jt}^{txt}x_{kt}^{txt}\beta_{jk}^{txt} \tag{14}$$

where x_t^{sea} captures seasonality, x_t^{imp} is the number of impressions, x_t^{pos} is the ad position, and x_t^{cpc} is the cost per click. The textual predictors representing the keyword search terms, x_t^{kwd}, and the ads, x_t^{ads}, are captured in the vector $x_t^{txt} = \left[\left(x_t^{kwd}\right)' \left(x_t^{ads}\right)'\right]'$. Finally, β represents parameters to be estimated.

5.3 Results

We first estimate our VANISH Poisson defined by Eqs. (13)–(14) on 3,203 daily observations and hold-out 168 daily observations for an out-of-sample analysis. The goal here is to demonstrate superior predictive validity. As such, we do not consider how to accommodate any potentially endogenous covariates. We compare our proposed VANISH Poisson model with 1,161 beta parameters with a baseline Poisson model with main effects only (67 beta parameters). The VANISH model incorporates

the same 67 main effects along with an additional 1,094 observed interaction effects (i.e., not all possible interactions are observed in the data). In-sample and out-of-sample fit statistics are reported in Tables 6 and 7. Estimates of the DIC show that VANISH model outperforms the baseline Poisson model with only main effects in-sample. Additionally, the VANISH Poisson outperforms a LASSO Poisson approach to including interactions. In terms of forecasting clicks based on 168 daily holdout observations we find that the VANISH model outperforms the base model as well as a LASSO and a MART Poisson model. These results show that incorporating interactions with our VANISH regularization provides superior predictive performance.

An interesting issue to consider is how to match the ad and keyword to generate more clicks. We use our VANISH results to match each of the 58 keyword with each of the ads to determine the keyword-ad pair that generates the highest number of clicks. We find that of the original 28 ads used by the firm, only 4 ads are retained as optimal when paired up with the 58 keywords. Of these ads, one is optimal for 28 keywords, the second is optimal for 13 keywords, the third is optimal for 13 keywords, and the fourth is optimal for 4 keywords. Based on these new keyword-ad pairs we estimate the counterfactual number of clicks in the holdout data. The results are presented in Fig. 1. The estimated number of clicks is 5.7 at the baseline rising 10.5% to 6.3 under our matching regime. Of course, a field test would be warranted to determine the actual increase in clicks. However, given that the 28 text ads are basically textual variations of the same ad it seems credible that indeed matching the keyword with an ad using a model is likely to increase the synergies between the keyword and the ad.

6 Summary and conclusion

Marketers increasingly face modeling situations where the number of independent variables is large and possibly approaching the number of observations. In this setting covariate selection and model estimation present significant challenges to usual methods of inference. These challenges are exacerbated when covariate interactions are of interest. The VANISH model (Radchenko and James 2010) addresses the issue of treatment of main and interactions terms in regularized linear regression models. The VANISH penalty ensures that the model follows the heredity principle (Nelder 1998). The degree of penalization on the interaction terms depends on whether the main effects are already present in the model. In this paper

Table 6 In-sample fit

Model	DIC[a]
Poisson (Main effects only)	−5,897.1
LASSO Poisson	−5,889.2
VANISH Poisson	−5,745.7

[a] Deviance information criterion based on Spiegelhalter et al. (2002)

Table 7 Out-of-sample fit

Model	MSE[a]	MAE[b]
Poisson (Main effects only)	51.78	2.86
Lasso Poisson	39.53	2.76
MART Poisson[c]	43.72	2.63
VANISH Poisson	28.70	2.54

[a] Mean squared error (using 100 forecasts, 168 Holdout Observations)

[b] Mean absolute error (using 100 forecasts, 168 Holdout Observations)

[c] 10,000 trees, 5 fold cross-validation

we generalize the VANISH model to accommodate non-linear response models of interest to marketing academics, including hazard, discrete choice, and count models. We also demonstrate how to accommodate unobserved heterogeneity in a VANISH regularization for generalized linear models. We adopt a Bayesian approach to inference which readily adapts to nonlinear response models and also accommodates the tuning parameters in the model hierarchy permitting simultaneous estimation of all model parameters. In a frequentist setting the tuning parameters of any regularization approach need to be inferred by using cross-validation techniques subsequent to parameter estimation.

In three empirical applications we demonstrate the usefulness of our proposed generalized VANISH approach. First, we apply a VANISH hazard model to the problem of modeling the time until a customer life-event in a customer relationship setting. Using a novel customer-level and predictor-rich dataset we show how to improve prediction of a life-event with our proposed model. We find that our proposed approach outperforms standard methods of forecasting life-events based on main effects only as well as other regularization approaches that consider interactions. Second, we apply a VANISH discrete choice model to the problem of predicting retweet behavior. We show that our generalized VANISH model ably predicts retweet

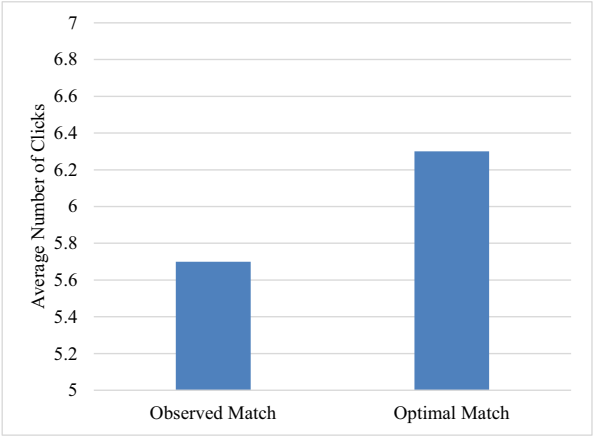

Fig. 1 Click estimates for observed vs. optimal keyword-ad match

behavior, out-performing alternative main effects and interaction models. We also introduce a VANISH model that accounts for unobserved heterogeneity using a mixture approach. We find that accounting for heterogeneity improves model performance. While the three segment VANISH model fits better in-sample is does not perform as well in holdout compared to a two segment VANISH model. In the third empirical application we consider consumer response to search text ads (i.e., the quantity of clicks on the ad) for a mobile app product. We model the count of ad clicks with a VANISH Poisson response model. We find that our generalized VANISH model results in superior in-sample and out-of-sample fit relative to alternative main effect and interaction models. Using the out-of-sample data we show that improving the ad-keyword match based on the model results in an improvement in the number of clicks.

In terms of limitations and future research the approaches presented in this paper do not consider the important problems of modeling unobserved heterogeneity with continuous distributions, state dependence, or more generally, dynamic behavior. Heterogeneous variable selection methods have received some attention in the literature. Adapting the VANISH prior to accommodate unobserved continuous heterogeneity in the response coefficients would be of interest to consider. One application of such an approach might be choice-based conjoint analysis where random coefficient choice models have become an industry standard and researchers are often interested in interaction effects. Likewise, variable selection in a dynamic setting is a nascent topic. Understanding variable selection in dynamic linear and non-linear models where interactions are of interest could also be of significant interest to marketers.

Appendix

We detail the steps in our VANISH regularization approach for the VANISH Hazard, VANISH Choice and VANISH Poisson models. For more information on the derivation of the full conditional distributions for the VANISH parameters (τ, ω, λ_1, λ_2) please see the Web Appendix.

Hazard model

1) Generate α and γ using a random-walk Metropolis-Hastings (MH) sampler based on the likelihood given by:

$$L = \prod_{i=1}^{n} \prod_{t=1}^{t_i} \Pr_i(t, x_{it})^{d_{it}} (1 - \Pr_i(t, x_{it}))^{1-d_{it}},$$

where

$$\Pr_i(t, x_{it}) = 1 - \frac{S_i(t, x_{it})}{S_i(t-1, x_{it})} = 1 - \exp\left(-\exp(x_{it}\beta) \int_{t-1}^{t} h_i(u)du\right)$$

where d_{it} is 1 if the life event occurs and zero otherwise.

2) Generateβ using a random-walk MH sampler based on the likelihood given by:

$$L = \prod_{i=1}^{n} \prod_{t=1}^{t_i} \Pr_i(t, x_{it})^{d_{it}} (1 - \Pr_i(t, x_{it}))^{1-d_{it}}$$

the VANISH prior given by:

$$\beta_j|...\propto \exp\left[-\frac{\left(\sum_{k=j+1}^{p} |\beta_{jk}| \right)^2}{2\tau_j^2} - \frac{\beta_j + \sum_{k:k\neq j} \beta_{jk}}{2\omega_j^2} \right].$$

3) Generate $\frac{1}{\tau^2}$

$$\frac{1}{\tau^2} = \gamma_j|...\sim InverseGaussian\left(\sqrt{\frac{\lambda_1^2}{\left(\sum_{k=j+1}^{p} |\beta_{jk}| \right)^2}}, \lambda_1^2 \right) I\left(\gamma_j > 0\right),$$

where $I(\cdot)$ is the indicator function.

4) Generate $\frac{1}{\omega^1}$

$$\frac{1}{\omega^1} = \varphi_j|...\sim InverseGaussian\left(\sqrt{\frac{\lambda_2^2}{(\beta_j)^2 + \sum_{k:k\neq j} (\beta_{jk})^2}}, \lambda_2^2 \right) I\left(\varphi_j > 0\right),$$

where $I(\cdot)$ is the indicator function.

5) Generate λ_1^2 and λ_2^2

$$\lambda_1^2|...\sim gamma\left(\frac{K}{2} + r, \frac{1}{2} \sum_{j=1}^{p} \tau_j^2 + s \right), \mathrm{K} = \#\text{main effects} + \#\text{interaction effects}$$

$$\lambda_2^2|...\sim gamma\left(\frac{H}{2} + r, \frac{1}{2} \sum_{j=1}^{p} \omega_j^2 + s \right), \mathrm{H} = 2*\#\text{main effects} + \#\text{interaction effects}$$

where $r = 1, s = 0.1$.

Choice model

1) Generate β^s using a random-walk MH sampler based on the likelihood given by (11) and (12) and the VANISH prior given by:

$$\beta_{jk}^s|...\propto \exp\left[-\frac{\left(\sum_{k=j+1}^{p}|\beta_{jk}^s|\right)^2}{2\left(\tau_j^s\right)^2}-\frac{\beta_j^s+\sum_{k:k\neq j}\beta_{jk}^s}{2\left(\omega_j^s\right)^s}\right].$$

2) Generate $\frac{1}{(\tau^s)^2}$

$$\frac{1}{(\tau^s)^s}=\gamma_j^s|...\sim InverseGaussian\left(\sqrt{\frac{\left(\lambda_1^s\right)^2}{\left(\sum_{k=j+1}^{p}|\beta_{jk}^s|\right)^2}},\left(\lambda_1^s\right)^2\right)I\left(\gamma_j^s>0\right),$$

where $I(\cdot)$is the indicator function.

3) Generate $\frac{1}{(\omega^s)^2}$

$$\frac{1}{(\omega^s)^2}=\varphi_j^s|...\sim InverseGaussian\left(\sqrt{\frac{\left(\lambda_2^s\right)^s}{\left(\beta_j^s\right)^2+\sum_{k:k\neq j}\left(\beta_{jk}^s\right)^2}},\left(\lambda_2^s\right)^s\right)I\left(\varphi_j^s>0\right),$$

where $I(\cdot)$ is the indicator function.

4) Generate $\left(\lambda_1^s\right)^2$ and $\left(\lambda_2^s\right)^2$

$$\left(\lambda_1^s\right)^2|...\sim gamma\left(\frac{K}{2}+r^{lam},\frac{1}{2}\sum_{j=1}^{p}\left(\tau_j^s\right)^2+s^{lam}\right),\ \mathrm{K}=\#\text{main effects}+\#\text{interaction effects}$$

$$\left(\lambda_2^s\right)^s|...\sim gamma\left(\frac{H}{2}+r^{lam},\frac{1}{2}\sum_{j=1}^{p}\left(\omega_j^s\right)^2+s^{lam}\right),\ \mathrm{H}=2*\#\text{main effects}+\#\text{interaction effects}$$

where $r^{lam}1$, $s^{lam}=0.1$.

5) Generate π

- $\pi|Z,\rho\sim Dir[(\tilde{\rho}_1...\tilde{\rho}_S)]$, where $\tilde{\rho}_s=\rho_s+\sum_{i=1}^{n}I(Z_i=s)$ with prior $\rho=(1,...,1)$.

6) Generate Z_i

- $Z_i \mid \theta, \mu, V, \pi \sim multinomial(1, [LR_1(\beta^i), ..., LR_S(\beta^i)])$,

where $LR_l(\beta^i) = \frac{\pi_l L(\beta^i)}{\sum_{s=1}^{S} \pi_s L(\beta^i)}$ and L is the likelihood given by (11) and (12).

7) Choose a permutation ξ_t^+ and relabel draws according to ξ_t^+

$$\xi_t^+ = \underset{\xi_t}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{s=1}^{S} p_{is}(\xi_t(\mu^t, V^t)) \log \left[\frac{p_{is}(\xi_t(\mu^t, V^t))}{\hat{q}_{is}^{t-1}} \right].$$

8) Set

$$\hat{Q}^t = \frac{t\hat{Q}^{t-1} + P\left(\xi_t^*(\mu^t, V^t)\right)}{t+1}.$$

Poisson model

$$\theta_t = x_t^{sea}\beta^{sea} + x_t^{imp}\beta^{imp} + x_t^{pos}\beta^{pos} + x_t^{cpc}\beta^{cpc} + \sum_j x_{jt}^{txt}\beta^{txt} + \sum_{j<k} x_{jt}^{txt}x_{kt}^{txt}\beta_{jk}^{txt}$$

1) Generate $[\beta^{sea}\beta^{imp}\beta^{pos}\beta^{cpc}]$ using a random-walk Metropolis-Hastings (MH) sampler based on the likelihood given by:

$$L \propto \prod_{t=1}^{n} \frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!}.$$

2) Generate β^{txt} using a random-walk MH sampler based on the likelihood given by:

$$L \propto \prod_{t=1}^{n} \frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!}$$

the VANISH prior given by:

$$\beta_{jk}|...\propto \exp\left[-\frac{\left(\sum_{k=j+1}^{p} |\beta_{jk}| \right)^2}{2\tau_j^2} - \frac{\beta_j + \sum_{k:k\neq j} \beta_{jk}}{2\omega_j^2} \right].$$

3) Generate $\frac{1}{\tau^2}$

$$\frac{1}{\tau^2} = \gamma_j | ... \sim InverseGaussian\left(\sqrt{\frac{\lambda_1^2}{\left(\sum_{k=j+1}^{p} |\beta_{jk}|\right)^2}}, \lambda_1^2\right) I\left(\gamma_j > 0\right),$$

where $I(\cdot)$ is the indicator function.

4) Generate $\frac{1}{\omega^1}$

$$\frac{1}{\omega^1} = \varphi_j | ... \sim InverseGaussian\left(\sqrt{\frac{\lambda_2^2}{\left(\beta_j\right)^2 + \sum_{k:k\neq j} \left(\beta_{jk}\right)^2}}, \lambda_2^2\right) I\left(\varphi_j > 0\right),$$

where $I(\cdot)$ is the indicator function.

5) Generate λ_1^2 and λ_2^2

$$\lambda_1^2 | ... \sim gamma\left(\frac{K}{2} + r, \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + s\right), \mathrm{K} = \#\text{main effects} + \#\text{interaction effects}$$

$$\lambda_2^2 | ... \sim gamma\left(\frac{H}{2} + r, \frac{1}{2}\sum_{j=1}^{p} \omega_j^2 + s\right), \mathrm{H} = 2*\#\text{main effects} + \#\text{interaction effects}$$

where $r = 1, s = 0.1$.

References

Agarwal, A.K.H. & Smith, M.S. (2011). Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets, *Journal of Marketing Research*, XLVIII (Dec), 1057–1073.

Allenby, G.M., Neeraj, A., & Ginter, J.L. (1998). On the Heterogeneity of Demand, *Journal of Marketing Research, 35*, 384–389.

Bakshy, E., Hofman, J.M., Mason, W.A., & Watts, D.J. (2011). Everyone's an influencer: Quantifying influence on Twitter. *Fourth ACM International Conference on Web Search and Data Mining*.

Bumbaca, F., Misra, S., & Rossi, P. (2017). Distributed Markov chain Monte Carlo for Bayesian hierarchical models. University of California, Irvine, *Working Paper*.

Cheng, J., Adamic, L., Dow, A., Kleinberg, J., & Leskovec, J. (2014). Can cascades be predicted? *Proc. 23rd International World Wide Web Conference*.

Ebbes, P., Papies, D., Van Heerde, H.J. The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science 30*(6),1115–1122

Ghose, A., & Yang, S. (2009). An empirical analysis of sponsored search in online advertising. *Management Science, 55*(10), 1605–1622.

Ghose, A., Ipeirotis, P. G., & Li, B. (2014). Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science, 60*(7), 1632–1654.

Gilbride, T. J., Allenby, G. M., & Brazell, J. D. (2006). Models for heterogeneous variable selection. *Journal of Marketing Research, 43*(3), 420–430.

Hong, L., Dan, O., & Davison, B. D. (2011). Predicting popular messages in Twitter. *WWW 2011*. Hyderabad, India.

Naik, P., Wedel, M., Bacon, L., Bodapati, A., Bradlow, E., Kamakura, W., Kreulen, J., Lenk, P., Madigan, D., & Montgomery, A. (2008). Challenges and opportunities in high-dimensional choice data analyses. *Marketing Letters, 19*(3), 201–213.

Nelder, J. A. (1998). The selection of terms in response-surface models—how strong is the weak-heredity principle? *The American Statistician, 52*(4), 315–318.

Park, T., & Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association, 103*, 681–686.

Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *The American Statistician, 44*(1), 26–30.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Petrovič, S., Osborne, M., & Lavenko, V. (2011). RT to win! Predicting message propagation in Twitter. *Association for the Advancement of Artificial Intelligence.*

Radchenko, P., & James, G. M. (2010). Variable selection using adaptive non-linear interaction structures in high dimensions. *Journal of the American Statistical Association, 105*, 1541–1553.

Rutz, O. J., Bucklin, R. E., & Sonnier. G. P. (2012). A latent instrumental variables approach to modeling keyword conversion in paid search advertising. *Journal of Marketing Research, XLIX*(Jun), 306–319.

Rutz, O. J., Sonnier, G. P., & Trusov, M. (2017). A new method to aid copy testing of paid search text advertisements. *Journal of Marketing Research, 54*(6), 885–900.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B, 64*, 583–639.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B, 62*(4), 795–809.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B, 58*(2), 267–288.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solution of ill-posed problems*. Washington: Winston & Sons.

Yoganarasimhan, H. (2018). Search Personalization using Machine Learning. Forthcoming at Management Science.

Zaman, T., Fox, E. B., & Bradlow, E. B. (2014). A Bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics, 8*(3), 1583–1611.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320.