AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-022-00465-5

ORIGINAL PAPER

Hedonic pricing modelling with unstructured predictors: an application to Italian Fashion Industry

Federico Crescenzi[1]

Received: 17 May 2021 / Accepted: 9 November 2022

Abstract
This study proposes a comparison of hedonic pricing models that use attributes obtained by featurizing text. We collected prices of items sold on the websites of five famous fashion producers in order to estimate hedonic pricing models that leverage the information contained in product descriptions. We mapped product descriptions to a high-dimensional feature space and compared predictive accuracy and variable selection properties of some statistical estimators that leverage sparse modelling, topic modelling and aggregated predictors, to test whether better predictive accuracy comes with an empirically consistent selection of attributes. We call this approach Hedonic Text-Regression modelling. Its novelty is that by using attributes obtained by text-mining of product descriptions, we obtain an estimate of the implicit price of the words contained therein. Empirically, all the proposed models outperformed the traditional hedonic pricing model in terms of predictive accuracy, while also providing consistent variable selection.

1 Introduction

The increasing role of e-commerce in consumers' purchasing behaviour gives researchers the opportunity to obtain detailed data on entire collections of products. These data are usually *unstructured*, and therefore, it lends itself to a wider range of methods of hedonic modelling that have not been addressed before.

In this study, we used product descriptions to estimate hedonic pricing models of fashion products sold on the online Italian stores of five famous brands. According to Archak et al. (2011), the primary weakness of hedonic models is the need to collect product features and in some circumstances to define measurement scales for them. As a matter of fact, websites lack of *structured* information on product attributes, which is mostly conveyed in form of product descriptions. These descriptions

✉ Federico Crescenzi
federico.crescenzi@unitus.it

[1] Department of Economics, Engineering, Society and Business, Universita degli Studi della Tuscia, Viterbo, Italy

Published online: 13 December 2022

are supposed to drive consumer's purchases so that they need to contain detailed information about features, usability, design and many other aspects.

Our modelling strategy started with text-mining of product descriptions to obtain their feature---or vector---space representation, which was used to compare predictive performances and variable selection properties of a class of hedonic pricing models that combine text-mining, sparse modelling and aggregated predictors.

From a methodological point of view, the main challenge posed by text data is high dimensionality. Text data are intrinsically high-dimensional, even when product descriptions are short. In this framework, traditional ordinary least squares estimation (OLS), if feasible, is likely to overfit the data, giving bad predictions of product prices. Also, not all the variables that can extracted from the text need to be significant predictors. It is actually more likely that very few words can significantly predict prices and so be regarded as real product attributes. For example, consider this product description: "A dappled print invades this lightweight blouse in silk georgette fabric. A style with a young, wild spirit that will sublimely complete casual looks with jeans or tailored trousers". The resulting hedonic model should estimate both the marginal price of materials (*silk*), the weight (*lightweight*) and the design (*young*, *wild*). According to Archak et al. (2011), these attributes are very difficult to include in a hedonic model that does not leverage unstructured predictors. From an economic point of view, to estimate a hedonic model using products descriptions is equivalent to estimating the implicit price of the words/attributes contained therein and, in turn, the hedonic value of the description.

As regards model selection and evaluation, traditional hedonic regression modelling relies on goodness of fit to select the best functional form of the model (Cassel and Mendelsohn 1985). Here, our model selection was performed by selecting the model that achieved the lowest prediction error. As pointed out by (Einav and Levin 2014), although this approach may sound obvious in other fields of research, it is not well-established in the field of empirical economics and applied econometrics, where big data applications have only recently been emerging.

To the best of our knowledge, the first (and so far only) attempt to incorporate unstructured attributes into a hedonic model is that of Nowak and Smith (2017) in real estate. The authors used text features to enlarge a given set of housing covariates (floor areas, number of bathrooms, etc.) and showed that this combination provided better predictions of prices. They compared the performances of two common penalized regression techniques, namely the LASSO (Tibshirani 1996) and the procedure proposed by Belloni et al. (2011). In this paper, we take the findings by Nowak and Smith (2017) further for at least for three main reasons. First, we do not have any structured set of covariates at our disposal. Second, we make use of ad hoc text-mining algorithms, namely latent semantic indexing (Deerwester et al. 1990) and latent Dirichlet allocation (Blei et al. 2003) to extract a set of features. Third, for the first time, we offer a comparative analysis of a wide range of estimators with different selection properties. Empirically, our application is concerned with the (Italian) fashion industry. To the best of our knowledge, no previous work has been conducted in this field.

The remainder of this paper is organized as follows. Section 2 provides details on traditional hedonic price modelling and how it can be adapted to the wide range of

approaches that text data offers. The data analysis is presented in Sects. 2-3. Conclusions and directions for further research are discussed in Sect. 4.

2 Hedonic price modelling

2.1 Hedonic text pricing model

Let us consider a standard Hedonic pricing model (HPM)

$$p_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{1}$$

where p_i denotes the price of item i, ϵ_i is a zero mean and finite variance error term and $\boldsymbol{x}_i \in \mathbb{R}^p$ a bundle of embodied attributes valued by some implicit or shadow prices embedded in the vector $\boldsymbol{\beta} \in \mathbb{R}^p$ (Baltas and Saridakis 2010). Compared to traditional HPMs, that models data as a collection of prices and attributes, $\{(p_i, \boldsymbol{x}_i)_{i=1}^n\}$, we model the collection $\{(p_i, \text{text}_i)_{i=1}^n\}$ where text_i is the textual description of product i. As text is unstructured, we map each text_i to a vector space representation in order to use it into a hedonic model.

Let $\mathcal{D}$ be a collection of D descriptions and let $\mathcal{V}$ be a set of V unique terms called the vocabulary of $\mathcal{D}$. In addition, let $\mathbf{W} \in \mathbb{R}^{D \times V}$ be a document-term matrix collecting the vector space representation of the descriptions. Entries or weights w_{ij} can be defined in many ways. For example, let tf_{ij} and df_j be the number of occurrences of word j in document i and the number of documents in the collection containing word j, respectively. Three popular ways of defining weights are known as term-frequency, term-presence and term-frequency-inverse-document-frequency (tf-idf). Under term frequency, we have $w_{ij} := \text{tf}_{ij}$. Weights under term presence are defined as $w_{ij} := 1_{\text{tf}_{ij}>0}$. Tf-idf is widely used in many text-mining applications, and it is defined as the ratio $w_{ij} := \text{tf}_{ij} \times \log^{-1}(D/\text{df}_j)$. Each of these weighting procedures were originally proposed by computer scientists in the field of text-mining, so that caution is needed when translating them into other areas of application. For example, tf-idf is widely used in text-mining as it downweights the effect of overly frequent words. However, in a regression framework, where the goal is to achieve a model with interpretable coefficients, a method like tf-idf may not be recommended. From an economic point of view, we suggest term presence as the most appropriate. In fact, the coefficient associated with the word has the simple and intuitive interpretation of the shift in the intercept associated with the presence of the word. On the contrary, tf-idf is a measure of the *linguistic* importance of a given word in the document and does not have a straightforward economic interpretation.

The hedonic pricing model can be then re-stated by setting $\boldsymbol{x}_i' := \boldsymbol{w}_i' \in \mathbb{R}^V$ to obtain the following model specification

$$p_i = \beta_0 + \beta_1 w_{i1} + \cdots + \beta_V w_{iV} + \epsilon_i \tag{2}$$

which we call the Hedonic text pricing model. The reason for this name is that estimating the coefficients in model 2 amounts to estimating the implicit value of every word in the description, and consequently the hedonic value of the description $\boldsymbol{\beta}'\boldsymbol{w}_i$.

Note that in this model, the order of words in the description has no importance, and therefore, it is sometimes referred to as a bag-of-words model.

Given the importance that words can have in explaining prices, a proper estimator of the vector of coefficients is fundamental. We consider three estimators of $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_V\}$ for this model. The first is of course the OLS estimator, which we use as benchmark. However, caution is necessary, first because V is usually an increasing function of D, so that V can be greater than D. In this case, we obtain infinitely many OLS solution for $\boldsymbol{\beta}$, one for each $v \in Kernel(W'W)$. Also, due to bias variance trade-off, this model is likely to give bad prediction of prices.

It is also a reasonable assumption that not all the words contained in the description are real product attributes, that is, have a nonzero shadow price. Consider again the introductory example: words like *looks*, and *spirit* would be weaker price determinants than *dappled, silk*, *georgette*, or *lightweight* in prices. For the first two, the shadow price is likely to be zero. Therefore, we assume that the true coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^V$ is sparse with support $\mathcal{S} = \text{support}(\boldsymbol{\beta}) \subset \{1, \ldots, V\}$. For this reason, instead of minimizing the residuals sum of squares, we solve the convex problem that minimizes the residual sum of squares plus a penalty term on the ℓ_1 norm of the vector $\boldsymbol{\beta}$, that is:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{D} \left(p_i - \beta_0 - \sum_{j=1}^{V} w_{ij}\beta_j \right)^2 + \lambda \sum_{i=1}^{V} |\beta_j| \tag{3}$$

The solution to this problem is known as LASSO (Tibshirani 1996). In 3, λ is a tuning parameter that is chosen via k fold cross-validation to minimize the prediction error. However, this choice may lead to a solution that includes too many false discoveries in $\hat{\mathcal{S}}$. This problem was considered in (Nowak and Smith 2017), where the authors used the false discovery rate (Benjamini and Hochberg 1995) to *screen* coefficients. Here, we consider an alternative estimator that enforces sparsity and puts an adaptive penalty on the sorted ℓ_1 norm of the coefficients to have control over the false discovery rate. This estimator is the solution to the following convex problem

$$\hat{\boldsymbol{\beta}}^{\text{slope}} = \arg\min_{\beta} \|\boldsymbol{p} - \mathbf{W}\boldsymbol{\beta}\|_2^2 + \lambda_1|\beta|_{(1)} + \lambda_2|\beta|_{(2)} + \ldots + \lambda_V|\beta|_{(V)} \tag{4}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_V$ and $\beta_{(1)} \geq |\beta|_{(2)} \geq \ldots \geq |\beta_{(V)}$. When the predictors are orthogonal and the variance of the error term σ^2 is known, the sequence of $\{\lambda_i\}$ is given by the Benjamini-Hochberg critical values. $\lambda_{BH}(i) = \Phi^{-1}(1 - q_i)$, where $q_i = i \times q/2V$ where $q \in (0, 1)$ and $\Phi^{-1}(\alpha)$ is the quantile of the standard normal distribution. When the value of σ^2 is unknown, there is some correlation between predictors the following procedure is used. First, the λ sequence is adjusted to $\lambda_G(1) = \lambda_{BH}(1)$ and for the rest $\lambda_G(i) = \lambda_{BH}(i)\sqrt{1 + w(i-1)\sum_{j<i} \lambda_G(j)^2}$ where the correction $w(k)$ is set equal to $(D - k - 1)^{-1}$. The procedure starts by setting the current subset of selected variables $\mathcal{S}_+$ equal to $\emptyset$, and then alternates iteratively between estimating σ^2 with a consistent estimator $\hat{\sigma}^2$, computing the solution $\hat{\beta}^{slope}$ with the sequence multiplied by $\hat{\sigma}^2$ and updating the set $\mathcal{S}_+$ until convergence. In the case of

non-orthogonal design, the correction is substituted with a Monte Carlo estimate. Further details on the implementation are discussed in (Bogdan et al. 2015).

2.2 Hedonic topic model

The bag-of-words model is not the only model available for text data. Alternative models for text data like latent semantic indexing (LSI) and latent Dirichlet allocation (LDA) represent documents as points in a lower-dimensional latent space defined by latent *concepts* or *topics*.

In LSI, in order to obtain the latent representation of documents as points in a topic space, we approximated the document-term matrix with a lower-rank (k) matrix $\mathbf{W}_k$ by computing its truncated singular value decomposition: $\mathbf{W}_k = \mathbf{U}_k \mathbf{\Sigma} \mathbf{V}'_k$. We have $\min_k \|\mathbf{W} - \mathbf{W}_k\|_2 = \sigma_k$, the k-th largest singular value representing the strength of the k-th topic inside the collection. To obtain a representation of documents in the k-dimensional topic space, we used matrix $\mathbf{U}_k$.

In LDA, documents are regarded as mixtures of topics, and topics are defined as distributions over the vocabulary. The data generating process is as follows:

1. Draw topics $\beta_k \sim \text{Dir}(\phi), k = 1, \ldots, K$
2. For each document
   (a) Draw topic proportions $\theta|\alpha \sim \text{Dir}(\alpha)$
   (b) for each word
      i. draw topic assignment $z_n|\theta \sim \text{Mult}(\theta)$
      ii. draw word $w_n|z_n;\beta \sim \text{Mult}(\beta_{z_n})$

In this model, the only variables that we observe are words, which we can gather into a DTM matrix. Other variables are latent. The joint distribution of both the observed and the latent variables is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{1:K}) \quad (5)$$

Inference under LDA consists of computing the posterior distribution of the latent variables conditional on the observed documents $p(\theta_{1:D}, z_{1:D}|w_{1:D})$. Algorithms for an approximation of this posterior distribution fall into two categories, namely sampling-based algorithms (Steyvers and Griffiths 2007) and variational methods (Blei et al. 2003; Wainwright and Jordan 2008). The choice is a trade-off between accuracy and computing time. In fact, in variational inference, the posterior is replaced with a distribution of the mean-field form $q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_n q(z_n|\phi_n)$ which is made close to the true posterior by obtaining the values $\{\gamma^*, \phi^*\}$ that minimize the Kullback--Leibler divergence between the two. This approach is faster than Gibbs sampling but of course it constitutes an approximation. In what follows, we use collapsed Gibbs sampling (Griffiths and Steyvers 2004) to integrate out the θ-s to

evaluate the posterior of the word-topic assignments given the observed documents $p(z|w)$. Once we have obtained the full conditional $p(z_d|z_{-d}, w)$, we can easily estimate θ and β by ratios of counts.

Both LSI and LDA are unsupervised in the sense that they only use the information in the DTM matrix. An easy way to obtain latent directions in the data in a supervised fashion is to look for a direction $\hat{z}_j$ that maximizes $\text{Corr}^2(\boldsymbol{p}, \mathbf{W}\boldsymbol{\alpha})\text{Var}(\boldsymbol{W\alpha})$ subject to $\|\boldsymbol{\alpha}\| = 1$ and $\boldsymbol{\alpha}'\mathbf{S}\hat{\boldsymbol{z}}_j = 0$, where $\mathbf{S}$ is the sample covariance matrix.

Suppose we want to predict the price of a t-shirt based on its description and that we learned a $k = 3$-dimensional topic representation to do so. Suppose also that we can interpret these topics as *quality of materials*, *appealing design* and *comfort*. We may then estimate the marginal effect of each topic on prices by letting

$$p_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_k z_{ik} + \epsilon_i \tag{6}$$

where z_{ik} is the representation of a given document in its *latent* topic space that is obtained by LSI or LDA. We refer to this model as a Hedonic topic regression model as it gives a Hedonic value to the topics in the documents.

2.3 Hedonic aggregated pricing model

Suppose that some attributes have the same shadow price. For example, we may assume that leather inserts on some apparels may have the same shadow price regardless of their type. Alternatively, suppose that the marginal effect of synthetic materials in prices is approximately the same regardless of these being *elastan* or *polyester*. Given that we are estimating a model based on absence or presence of some words relating to attributes, it seems reasonable to assume that attributes that belong to the same *group* may share very similar shadow prices.

Let $(\mathcal{G}_i)_{i=1}^G$ be a partition of indices $\{1, 2, \ldots, V\}$, where there exist a one-to-one relation between this set and each word in the vocabulary $\mathcal{V}$. Under this assumption, the model in 2 can be re-formulated as

$$p_i = \beta_0 + \sum_{g=1}^{G} \beta_g \sum_{j=1}^{V} w_{ij} \mathbb{1}_{\{j \in \mathcal{G}_g\}} + \epsilon_i. \tag{7}$$

Under this model, covariates that belong to the same set $\mathcal{G}_g = \{j_1, \ldots, j_{|\mathcal{G}_g|}\}$ have the same coefficient β_g. This approach substitutes the original set of covariates with a new one each which is the sum of the weights of the words in the group. Under a term presence weighting, each coefficient gives the shadow price of an additional count of one of the words in the group.

Park et al. (2007) proved that under some conditions on the sample covariance structure of predictors, identifying and consolidating predictors into groups reduces prediction error. Without loss of generality, assume that $\mathbf{W}'\mathbf{W} = \mathbf{I}$ and let $\tilde{\boldsymbol{\beta}} = (\hat{\beta}_a, \ldots, \hat{\beta}_a)' \in \mathbb{R}^V$ be a vector such that $\hat{\beta}_a$ is the OLS coefficient when p is regressed onto the sum of the predictors. Then, $\mathbb{E}_{p|W}[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2] < \mathbb{E}_{p|\mathbf{W}}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2]$ if and only if $\rho > 1 - (V-1)\sigma^2 / \sum_v (\beta_v - \bar{\beta})$, where ρ is the correlation of any given pair of predictors. This means that if the true coefficients of predictors are similar

then the range of ρ to improve the fit is large. The authors provide a two-step procedure to find significant groups of covariates. The first step is applying hierarchical clustering to the design matrix and averaging within groups at each level of the hierarchy. The second step is to use cross-validation to find the optimal level of hierarchy by means of the LASSO.

To select groups of words, we refer to the algorithm used in Park et al. (2007) but with a slight modification. By leveraging the dual interpretation of the vector space of text, that is words in document space, we form clusters of words using cosine similarity. Cosine similarity is widely used in text-mining literature for measuring similarities between words for text clustering (Berry and Kogan 2010; Berry and Castellanos 2004; Aggarwal 2018). Cosine similarity between two word vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^D$ is defined as $\boldsymbol{v}_1' \boldsymbol{v}_2 / (\|\boldsymbol{v}_1\| \|\boldsymbol{v}_2\|)$.

An alternative method of forming aggregated predictors is to assume a tree-guided aggregation (Yan and Bien 2020). Let $\mathcal{T}$ be a tree with leaves $\{1, \ldots, V\}$. To aggregate the β-s into branches, a parameter γ_u is assigned to each node u in $\mathcal{T}$. Then, each coefficient is set such that $\beta_j = \sum_{u \in ancestor(j) \bigcup \{j\}} \gamma_u$ and we seek the solution of the following problem: $\min_{\beta,\gamma} 1/(2D) \|\boldsymbol{p} - \mathbf{W}\boldsymbol{\beta}\| + \lambda(\alpha \|\gamma_{-root}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_1$ such that $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\gamma}$ where $\mathbf{A} \in \{0,1\}^{V \times |\mathcal{T}|}$ and $A_{jk} = 1$ if node u_k is an ancestor of coefficient j. The difference from the previous approach is that the penalties on γ and β simultaneously induce aggregation of coefficients and variable selection. For $\alpha = 1$, this is equivalent to a LASSO problem in γ, while for $\alpha = 1$, this is equivalent to a LASSO problem in β. For values $\alpha \in (0,1)$, this can be solved as generalized LASSO problem (Tibshirani and Taylor 2011).

3 Data

3.1 Data collection

Cachon and Swinney (2011) define four different systems by which firms operate in the fashion market: *traditional*, *enhanced-design*, *quick-response* and *fast-fashion*. The distinction is straightforward. Traditional firms have long production lead times and standard product design abilities. This system closely resembles a newsvendor model. Enhanced-design (ED) firms rely on enhanced design to increase consumer willingness to pay but avoid the kind of radical supply chain necessary to achieve lead time reduction. Quick-response firms do not employ enhanced design capabilities, but have significantly shorter production lead times. Fast-fashion firms exploit both quick response and enhanced design capabilities.

For this application, we collected publicly available data from the Italian websites of five fashion brands: Zara (Z), H &M, Pinko (P), Patrizia Pepe (PP) and Elisabetta Franchi (EF). We limited our attention to the categories of women's trousers (pants) and dresses. We collected the price (in euros) and (text) description of each item from the brand websites. This is an important point, as descriptions serve as a

Table 1 Data after pre-processing

Category	Sample size	Number of features
Trousers	692	302
Dresses	814	309

marketing purpose, the set of features that we can obtain by their mining is reasonably containing a good set of attributes that we can use as predictors[1].

Hedonic pricing models such as those mentioned above are usually also specified in terms of a time dummy variable. However, while we were collecting the data, we observed that prices almost always remained unchanged. What we noticed instead was that products have a life cycle: they come out and are then dropped without a price change. It was rare to observe discounts but in any case, our crawler was designed to collect original, not discounted, prices. Regarding seasonality, in order to derive meaningful hedonic models, our crawling routine ran on a weekly basis from October to January before the beginning of the sales period[2]. In fact, in this period, a downshift in the level of prices would result in an underestimation of the marginal effects.

Finally, the selection of brands was made with the purpose to resemble the market as described in Cachon and Swinney (2011). Therefore, what follows should not be erroneously interpreted as a classification problem of items into these segments. As stated, our purpose was to derive the hedonic models described in the sections above.

3.2 Pre-processing

Since each crawling routine makes a copy of the whole website, the resulting data set contains big overlaps for the many data collection dates. Records that had the same price and description were therefore discarded, keeping just one record for each. Pre-processing is usually useful to have better vector space representation of text data. We used standard pre-processing steps; we removed punctuation, accents, special characters, numbers and stopwords[3]. The final step was stemming, namely reducing each word to its root. This step need not always produce meaningful words in general.

Table 1 reports the number of observations and the number of text features after the pre-processing steps.

Note that we are not left with much sample size. While this is somehow obvious since we are bounded by the effective stock of items shown on the websites

[1] There may be sections of websites with regulation rather than marketing information on products. By scraping and mining this data, the set of significant attributes that we obtain is more likely a consequence of obvious causality rather than actual price determinants.

[2] In Italy, depending on the region, usually starts at mid-January.

[3] Stopwords are words that do not provide any semantic meaning. The descriptions that we use are Italian descriptions.

for a given collection, to keep on scraping data is not a good choice for the reasons illustrated above (sales period approaching and new collections to come). However, the methods proposed in this paper are robust to small sample size. For example, LASSO and SLOPE are designed to work even when the number of covariates is much greater than the sample size.

3.3 Pricing strategies

We compared the in-sample and out-of-sample predictive accuracy of the following fits:

$$[\text{OLS}]: \hat{p}_i = \boldsymbol{w}_i' \hat{\boldsymbol{\beta}}^{ols} \tag{8}$$

$$[\text{LASSO}]: \hat{p}_i = \boldsymbol{w}_i' \hat{\boldsymbol{\beta}}^{lasso} \tag{9}$$

$$[\text{SLOPE}]: \hat{p}_i = \boldsymbol{w}_i' \hat{\boldsymbol{\beta}}^{slope} \tag{10}$$

$$[\text{LSI}]: \hat{p}_i = \boldsymbol{u}_i' \hat{\boldsymbol{\beta}}^{ols} \tag{11}$$

$$[\text{LDA}]: \hat{p}_i = \boldsymbol{z}_i' \hat{\boldsymbol{\beta}}^{ols} \tag{12}$$

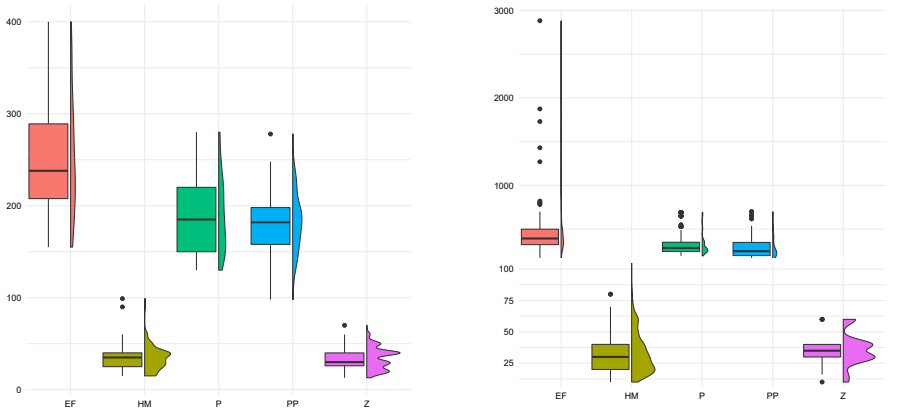
$$[\text{PLS}]: \hat{p}_i = \boldsymbol{w}_i' \hat{\boldsymbol{\beta}}^{pls} \tag{13}$$

$$[\text{AP}_G]: \hat{p}_i = \boldsymbol{w}_i^G \hat{\boldsymbol{\beta}}_G^{ols} \tag{14}$$

$$[\text{AP}_{\mathcal{T}}]: \hat{p}_i = \boldsymbol{w}_i \hat{\boldsymbol{\beta}}^{\mathcal{T}} \tag{15}$$

For the fits in 11 and 12, the number of latent topics k needed specification. In order to select the model corresponding to the best values of $\{\lambda, k\}$, we used k as an additional tuning parameter chosen by cross-validation, that is $(k^*, \lambda^*) = \underset{k,\lambda}{\text{argmin}}\, \text{CVErr}(\hat{f}_{\lambda,k})$. This procedure over k is akin to that used in principal component regression which we tried to improve in predictive accuracy by further regularization.

Interestingly, given that $\mathbf{U}_k'\mathbf{U}_k = \mathbf{I}$, we obtained a closed form solution to solution to Equation 3 so that $\hat{\beta}_j$ is equal to $\boldsymbol{u}_j'\boldsymbol{y} + \lambda$ if $\boldsymbol{u}_j'\boldsymbol{y} < -\lambda$, it is equal to 0 if $|\boldsymbol{u}_j'\boldsymbol{y}| < \lambda$ and it is equal to $\boldsymbol{u}_j'\boldsymbol{y} - \lambda$ if $\boldsymbol{u}_j'\boldsymbol{y} > \lambda$. Therefore, if the strength of the linear dependence between $\boldsymbol{u}_j$ and $\boldsymbol{y}$ measured by its OLS coefficient $\hat{\boldsymbol{\beta}}^{OLS} = \boldsymbol{u}_j'\boldsymbol{y}$ exceeded the value of λ, then the j-th topic was included in the model with a shrunk coefficient.



(i) Trousers (ii) Dresses

Fig. 1 (i) Trousers (ii) Dresses. Price distributions

Table 2 Trousers: price distribution descriptive statistics

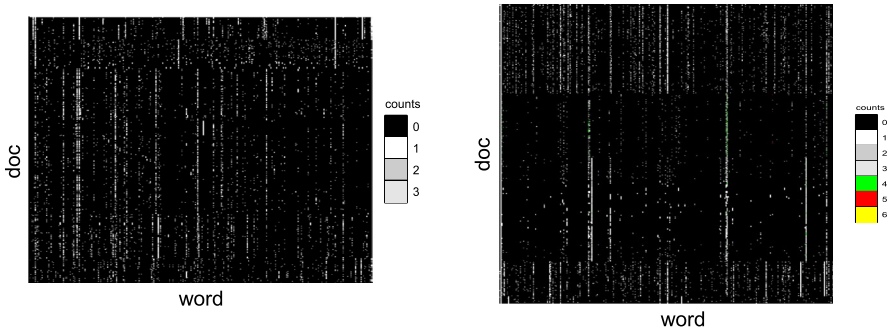
Retailer	*n*	Min	1st Q.	Mean	Med.	3rd Q.	Max
EF	58	155	207.75	250.65	238	289	400
HM	114	14.99	24.99	34.41	34.99	39.99	99
P	76	130	150	188.81	185	220	280
PP	65	98	158	182.06	182	198	278
Z	379	12.95	25.95	33.92	29.95	39.95	69.95

The fit in Equation 12 used LDA to estimate the topic structure. We obtained this estimate using collapsed Gibbs sampling, assuming symmetric Dirichlet priors. This required the additional parameters α and ϕ to be specified. These parameters correspond to sparse, uniform or bumped Dirichlet distributions over the topic and word simplexes. For example, to higher values of ϕ, the number of topics to describe the dataset is expected to decrease. As with LSI, we select these parameters via cross-validation.

The fit in 13 which uses partial least squares may also be understood as a topic hedonic model. The difference with respect to the other topic models is that this fit is meant for a joint modelling of both topics and prices by partial least squares.

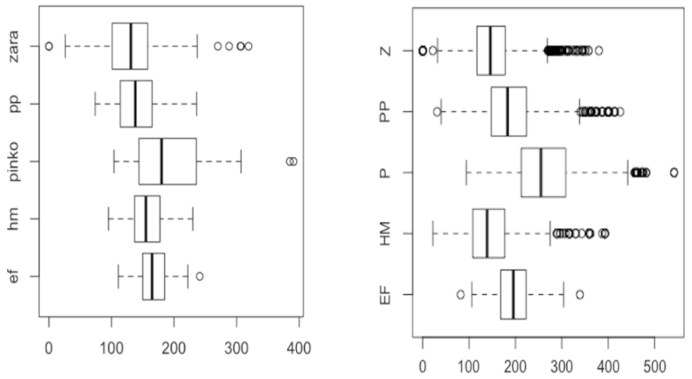
4 Results

Trousers Figure 1i shows that two couples of fashion retailers, (ZZ and HM) and (P and PP) had similar, nearly overlapping, price distributions. Also, ZZ and HM showed peaked distributions as their pricing strategy is to concentrate prices into different levels. For more details on prices, see Table 2.



(i) Trousers (ii) Dresses

Fig. 2 (i) Trousers (ii) Dresses. Document-term matrices



(i) Trousers (ii) Dresses

Fig. 3 (i) Trousers (ii) Dresses. Word counts distributions

Figure 2i shows the document-term matrix for the descriptions of trousers. Counts were really sparse and most features occurred only once across documents. The most frequent words were associated with the fit (tight, low waist) and the design (classic, bell-shaped). Figure 3i shows the distribution of word counts in product descriptions of each retailer. A few differences stood out between retailers; the distributions of E, H and PP were nearly equal, whereas Z showed greater variability. A mean of 146 words was used to describe trousers.

Figure 4i shows pairwise word correlations. We notice that while some groups of words exhibit high correlations, these are very low in general.

Table 3 reports in-sample goodness of fit and predictive accuracy of the models that we tested. In the first part of the table, the *base* fit is $\hat{p}_i = \bar{p} = n^{-1} \sum_i p_i$; in the second part, it is the least squares fit allowing for brand effects taking EF as a baseline. For this model, the marginal effects for Z and HM were nearly the same (-216.731 and -216.244, respectively), while the marginal effects for PP and P were

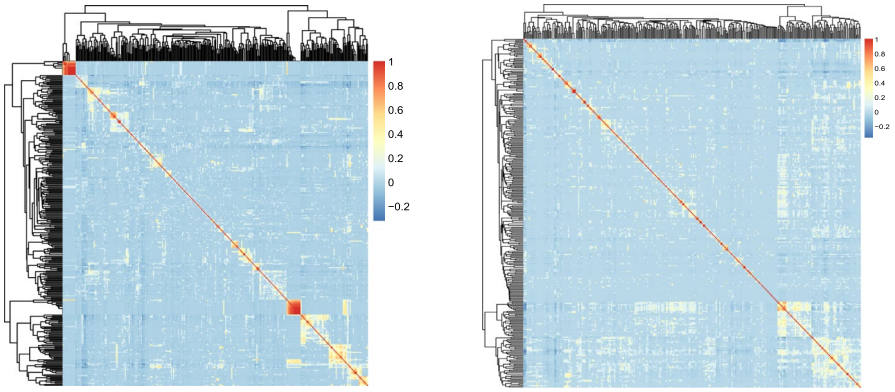


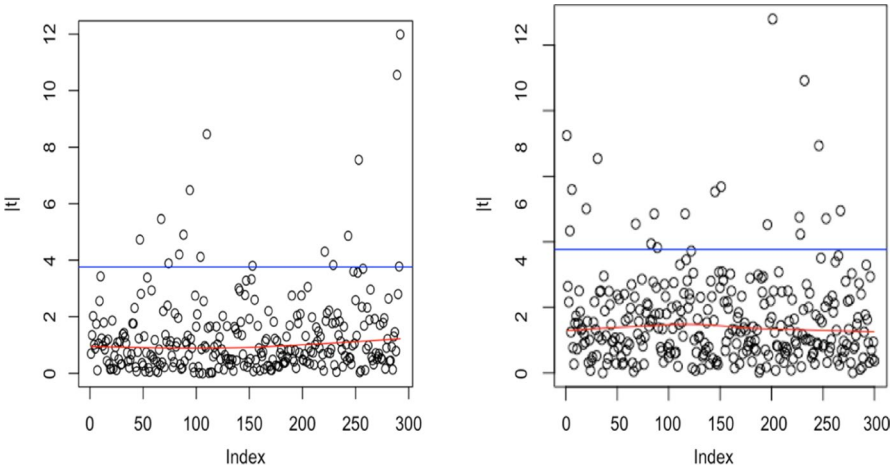
Fig. 4 (i) Trousers (ii) Dresses. Word correlations

Table 3 Results of hedonic text pricing models for trousers

	No brand				Brand			
Fit	$\lvert\hat{\mathcal{S}}\rvert$	adj-R^2	$RMSE_{cv}$	RSE	$\lvert\hat{\mathcal{S}}\rvert$	adj-R^2	$RMSE_{cv}$	RSE
BASE	1	–	83.339	1	5	0.897	26.707	0.103
Text								
OLS	302	0.947	70.056	0.701	306	0.947	60.205	0.552
Sparse								
LASSO	82	0.901	35.428	0.181	122	0.928	20.948	0.063
SLOPE	85	0.913	34.193	0.168	59	0.906	22.055	0.070
Topic								
LSI-LASSO	96	0.884	32.167	0.149	138	0.961	21.221	0.065
LDA-LASSO	58	0.855	33.529	0.162	84	0.936	23.002	0.076
PLS	4	0.902	27.629	0.110	8	0.902	15.653	0.035
Aggregated								
AP_{10}	10	0.698	46.753	0.315	10	0.919	25.919	0.097
AP_{20}	20	0.793	39.212	0.221	20	0.922	24.535	0.087
$AP_{\mathcal{T}}$	68	0.932	31.637	0.144	186	0.945	21.468	0.066

-68.594 and -61.839, respectively. This suggests similar market collocations with respect to the reference brand EF.

Brand was clearly an important determinant of price, as the estimated prediction error was one of the lowest at 26.707. Interestingly, using OLS, we obtained the highest value of R^2 and the worst predictive accuracy of all the proposed fits. In addition, while including text features enhanced the predictive accuracy from BASE to OLS, it impaired it when combined with the brands.



(i) Trousers (ii) Dresses

Fig. 5 (i) Trousers (ii) Dresses. Red line is a lowest fit of data. The blue line is the Bonferroni threshold

Table 4 Trousers: coefficient estimates exceeding Bonferroni threshold

	Estimate	Std. error	t value
(Intercept)	332.643	25.922	12.832
Without	136.329	28.812	4.372
Coutur	-271.609	49.743	-5.46
Virgin	201.003	51.72	3.886
Nappa	161.357	38.385	4.204
Synthetic	-230.928	47.11	-4.902
Iper	-295.894	45.725	-4.119
Profile	-196.582	47.725	-4.119
Zip	-376.026	44.402	-8.469
Palazzo	76.026	19.988	3.804
Elastan	191.844	44.567	4.305
Logo	-126.307	32.981	-3.83
Poliester	-134.306	27.613	-4.864
Invisible	-262.211	34.691	-7.558
Retailerhm	-283.748	26.882	-10.555
Retailerpp	-104.164	27.567	-3.779
Retailerzara	-298.693	24.91	-11.991

An F-test rejected the hypothesis that the effect of the words is insignificant at level $\alpha = 0.05$. We also tested for statistical significance using Bonferroni correction for $\alpha = 0.05$. Figure 5i shows the distribution of absolute t-statistics for linear regression of price on brand fixed effects and text features. Coefficients that fall above the threshold are reported in Table 4.

It is interesting to see the estimates corresponding to the words that denote products' materials. As expected, the presence of synthetic materials (*synthetic*, *polieester*) leads on average to lower prices, while high-quality materials like virgin wool (*virgin*) leads to higher prices. Words relating trouser details and design also led to higher prices. This was the case of *palazzo* style trousers and those with attributes denoting minimal design (effect of word *without*).

As expected, we obtained lower prediction errors when we applied regularization. Using LASSO, the CV estimate of prediction error fell from 70.056 to 35.428 (about 50% lower) with 82 nonzero coefficients and from 60.205 to 20.948 (66% lower---with 122 nonzero coefficients).

The hedonic model estimated using SLOPE provided very similar results to LASSO when only the text features were considered. The estimate of prediction error was slightly better and the two models selected approximately the same number of variables. The situation was different when we allowed for brand effects. While LASSO kept more features in the model than SLOPE, the selection made by SLOPE was stricter. This may be because SLOPE has better control of false discoveries, while LASSO tends to select too many irrelevant as it seeks for minimization of prediction error. In fact, the estimate of the prediction error was slightly lower than that of SLOPE.

Topic hedonic models based on the topic content of the descriptions showed satisfactory predictive accuracy, but the number of topics selected was too large to interpret. Nevertheless, we obtained the best results using PLS.

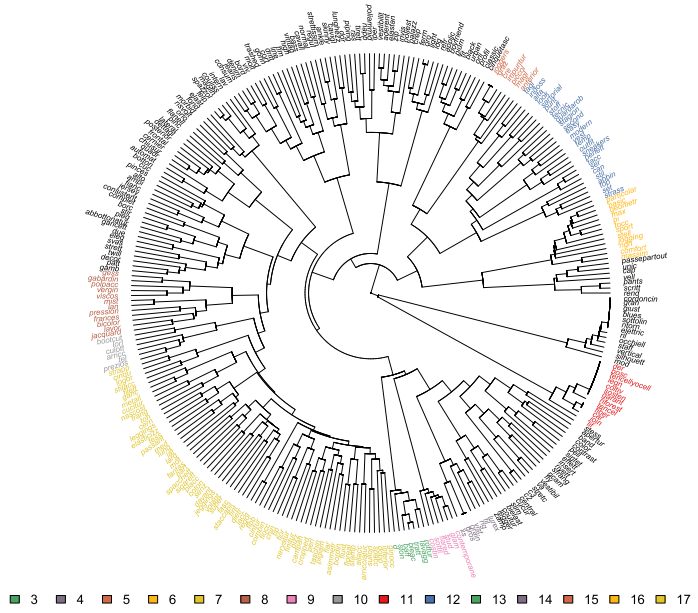
The predictive accuracy of the models that used aggregated predictors was the worst of all. Still, our results empirically confirm the results of Park et al. (2007). The values $g = 10, 20$ were selected arbitrarily to avoid a clustering structure too dependent on brand effects that could have prevailed for values close to the effective number of brands available. Regarding interpretation of the results, we noticed that some superwords[4] have a nice practical interpretation. For example, in Figure 6i, Group 1 gathers the effect of tencel (*tencel, tencelyoce, fiber*) and Group 2 gathers words recalling the effect of the overall outfit (*abbin*, outfit, wardrobe) which express the value of the item when paired with others. Group 3 contains features related to maintenance (*rottur, lavaggio, trattamento*). Group 8 collects words regarding trousers with skinny fit.

The tree guided aggregation proposed by Yan and Bien (2020) showed good predictive accuracy but kept too many groups of single words in the model when it is used in conjunction with brand effects. On the contrary, the performance was satisfactory when we considered the no-brand scenario, as it was better than that of the LASSO and SLOPE. Also, the cv-estimate of α was equal to 1, which means that the grouping property exerts all the weight.

Dresses The same analysis was repeated for dresses. We had 814 observations and 309 textual features from descriptions. The document-term matrix obtained by pre-processing the description is displayed in Fig. 2ii. The most frequent words

[4] we use the notion of superwords to refer to group of words. It resembles the notion of supergenes introduced by Park et al. (2007)

(i) $g = 10$ *groups*



(ii) $g = 20$ *groups*

Fig. 6 (i) $g = 10$ *groups* (ii) $g = 20$ *groups*. Hierarchical clustering of word vectors for trousers DTM

Table 5 Dresses: price distribution descriptive statistics

Retailer	*n*	Min	1st Q.	Mean	Med.	3rd Q.	Max
EF	95	171	323	493.63	393	499	2882
HM	242	9.99	19.99	32.93	29.99	39.99	129
P	281	195	245	320.44	285	350	690
PP	176	148	198	293.76	248	348	698
Z	20	9.99	29.95	36.05	34.95	39.95	59.95

Table 6 Results of hedonic text pricing models for dresses category

	No brand				Brand			
Fit	$\lvert\hat{\mathcal{S}}\rvert$	adjR^2	RMSE$_{cv}$	RSE	$\lvert\hat{\mathcal{S}}\rvert$	adjR^2	RMSE$_{cv}$	RSE
BASE	1	–	219.709	1	4	0.499	150.27	0.468
Text								
OLS	309	0.923	236.437	1.158	313	0.928	199.851	0.827
Sparse								
LASSO	99	0.9	113.256	0.266	117	0.905	112.094	0.26
SLOPE	87	0.897	108.935	0.246	68	0.904	102.127	0.216
Topic								
LSI-LASSO	154	0.863	119.996	0.293	121	0.829	112.094	0.26
LDA-LASSO	9	0.554	148.203	0.455	32	0.652	132.069	0.361
PLS	14	0.785	101.994	0.216	18	0.933	53.367	0.059
Aggregated								
AP$_{10}$	10	0.433	166.792	0.576	10	0.521	154.009	0.491
AP$_{20}$	20	0.548	156.697	0.509	20	0.630	141.643	0.416
AP$_{\mathcal{T}}$	90	0.841	118.357	0.290	90	0.852	112.197	0.260

in the collection concerned materials: poliester, viscose, elastan, cotton and wool. Figure 3ii shows the distribution of word counts in the description of each retailer. As observed for trousers, Z and HM showed similar distributions, did PP and EF. Retailer P used the most words to describe its products. Figure 4ii shows the pairwise correlations between words in the collection. While some groups showed high correlations, in general words were not highly correlated.

Figure 1ii shows the price distribution of each retailer. It is clear that prices for fast fashion retailers like HM and Z have much lower average and median prices than their competitors. More details are given in Table 5.

The results in Table 6 offer insights similar to those observed for trousers[5] although here brands explained a much lower proportion of the total variability. The marginal effects for each retailer are HM -460.704, P -173.187, PP -199.870 and Z -457.580.

[5] we excluded the dresses priced over 1000 euros for brand EF as of Figure 1ii.

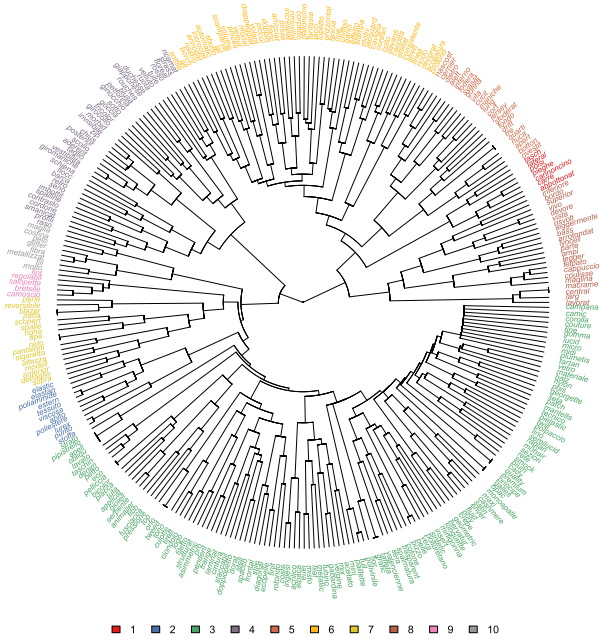
Table 7 Dresses: coefficient estimates exceeding Bonferroni threshold

	Estimate	Std.error	t value
Intercept	426.707	59.85	7.13
Acetate	-365.793	50.423	-7.254
Gathered	132.114	26.365	5.011
Blouse	-346.968	52.988	-6.548
Eco	446.737	113.392	3.94
Elastan	-23.531	4.845	-4.857
Outer	67.311	17.6	3.824
Georgette	91.535	18.842	4.858
Light	168.423	30.44	5.533
Shiny	166.482	29.282	5.685
Pettorina	528.386	116.757	4.526
Feather	579.293	53.632	10.801
Embroidered	166.307	34.876	4.769
Stiped	-92.342	21.812	-4.234
Red	789.002	88.476	8.918
Evening	210.189	30.283	6.941
Sleeveless	-183.753	38.99	-4.713
Strass	207.702	41.938	4.953
RetailerHM	-365.793	50.423	-7.254
RetailerZ	-334.407	77.141	-4.335

The multiple testing procedure using the Bonferroni correction produced some interesting findings. As reported in Table 7, words denoting the low-quality materials *acetate* and *elastan* had coefficients with a negative sign. Dresses made with ecofeatures (*eco*) or made with *georgette* were given higher prices on average. Also lightweight dresses were given higher prices. Among product details, we notice that colour *red* had a positive impact on prices, as well as *strass* and embroidered finishes. Remarkably, the coefficient for evening dresses (*sera*) was given positive sign.

Dimension reduction via LSI or LDA leads to higher prediction errors than those obtained using sparse modelling. In particular, the prediction error we obtained by dimension reduction using LSI was approximately 5% higher than that obtained using LASSO and 9% higher than that obtained with SLOPE in the no brand specification. Brand specification improved the overall predictive accuracy, but did not change the ranking of the methods. As observed for trousers, we obtained the best results using a topic hedonic model estimated with PLS. In particular, PLS with no brand effects also performed better than other fits that included brand specification.

Compared to other models, hedonic regression with aggregated predictors again gave the worst fit. This was somehow expected since we forced words into a small number of clusters. However, the predictive accuracy of this fit was better than that obtained with ordinary least squares, even if the model explained less variance in



(i) $g = 10$ *groups*

(ii) $g = 20$ *groups*

Fig. 7 (i) $g = 10$ *groups* (ii) $g = 20$ *groups*. Hierarchical clustering of word vectors for dresses DTM

data. In particular, prediction error was on average 27% lower with brand specification and 32% lower without. This is consistent with the theoretical results of Park et al. (2007). Figure 7i shows that group 2 collected words associated with synthetic materials (*elastic, elastan, poliammide, viscosa*). This means that words belonging to this groups had the same marginal effect on price. Another example: group 9 is salopettes, so attributes of this subcategory contributed equally to prices. In Fig. 7ii, group 20 denoted a superword collecting attributes such as fur coat and leather finishes (*leather, snake*). Group 14 gathered clearly high quality attributes denoted by the words *mohair, premium, cashmere, galles*.

5 Conclusions

In this study, we built hedonic pricing models of Italian fashion products using the internet as a source of data and attributes obtained by text-mining product descriptions. We tested and compared the predictive performance and variable selection properties of a series of models that use sparse estimators, dimension reduction, grouping of predictors and a combination of the last two with the first. We tested the models, either including or excluding brand effects, and as expected, we found that brand effects captured the most significant part of the underlying signal. Also, we may see brand effects as an estimate of the average market price for each of the systems (enhanced, traditional, quick-response, fast-fashion) discussed by Cachon and Swinney (2011).

Empirically, all the models we suggested outperformed the traditional hedonic pricing model. However, our results showed that it is not straightforward to balance predictive accuracy with interpretability, as one usually comes at the expense of the other.

Our approach was different from the previous attempt by (Nowak and Smith 2017), since we did not investigate whether product descriptions could be used to improve predictive accuracy when used in combination with other control variables. In this framework, structured information is missing or unreliable, making it necessary to build on text data alone.

An appealing feature of working with product descriptions is that the set of predictors that can include attributes that may not be directly observable (for example, aspects of fit or design). The set is therefore much richer than one obtained by just looking for structured covariates. As a consequence, this richer specification is likely to mitigate the bias due to omitted variables, a common issue in hedonic models. On the contrary, a drawback is that product descriptions can contain many noisy variables, but interestingly, the variable selection procedures that we tried proved to be robust, giving coefficients with the expected signs. For example, the colour red, furs, leather and embroidery were all attributes that bring in higher prices for dresses, including evening dresses, while low quality materials like elastan reflect to lower prices. We also found this in our grouping procedure, where the groups were inherently consistent and had the expected effect on the prices.

Nevertheless, a study by Archak et al. (2011) does not recommend working with descriptions since the text is too static and says little about the characteristics of

the goods. On the contrary, our findings provide empirical evidence that this is not necessarily true. In fact, when compared to a baseline model where the price is regressed solely on brand effects, the set of attributes we obtained from descriptions dramatically improved the predictive performance of the model. This is an interesting result, since clothes are known to be heavily influenced by brand names, as we also observed empirically in our dataset. So while it is certainly possible to estimate a model with text features only, one can expect these to become insignificant when the brand effect is considered.

As of the major implications of this study, apart from the recent applications in the field of official statistics (Cavallo 2017, 2018), this framework is also profitable for many companies that operate via web or app. For example, in 2017, the biggest community-powered shopping application in Japan, Mercari, launched a Kaggle competition to develop pricing algorithms to suggest product prices to sellers based on a text description of the product[6]. Of course, the ultimate interest for a company like Mercari is to achieve the best predictive accuracy, rather than interpretable and consistent results. Thus, a company like Mercari would probably opt for a topic model based on PLS. Nevertheless, this framework may also be of practical use for brands themselves. Consider the launch of a new product: the brand marketing office could just type in a description of what it wants to produce and the model would return a price; or the office may be interested in predicting the price of its competitors. This is certainly a great advantage for pricing strategies, especially when it comes to deriving the marginal effect of attributes used by competitors. A model seeking interpretability over prediction error can also be of interest from a consumer perspective. In fact, consumers can be made aware of the marginal mark-ups of a brand or given attributes like a certain colour, finish or design. In this case, a sparse hedonic model estimated with SLOPE is probably preferable, since the selection of variables aims to control false discoveries as much as possible.

A potential limit of our results is that penalization/grouping of coefficients was done by considering the whole vector of covariates including brand effects. One may argue that by not penalizing/grouping brand coefficients, we may be more able to identify which tokens are relevant to price, above and beyond the implicit brand-specific effects common to all items sold by a brand. While an l1 penalization may be accounted for solely by the subset of tokens, this was not addressed in other methods discussed here, where selection/grouping properties may not hold.

References

Aggarwal, C.C.: Machine learning for text. Springer, Newyork (2018)

Archak, N., Ghose, A., Ipeirotis, P.G.: Deriving the pricing power of product features by mining consumer reviews. Manag. Sci. **57**(8), 1485–1509 (2011)

Baltas, G., Saridakis, C.: Measuring brand equity in the car market: a hedonic price analysis. J. Oper. Res. Soc. **61**(2), 284–293 (2010)

[6] The competition rewarded the best three pricing algorithms in terms of root-mean-squared logarithmic error with a monetary prize of 60 k, 30 k and 10 k dollars, respectively

Belloni, A., Chernozhukov, V., Wang, L.: Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika **98**(4), 791–806 (2011)
Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc.: Ser. B **57**(1), 289–300 (1995)
Berry, M., Kogan, J.: Text mining: applications and theory. Wiley, Newjersey (2010)
Berry, M.W., Castellanos, M.: Survey of text mining. Comput. Rev. **45**(9), 548 (2004)
Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., Candès, E.J.: Slope–adaptive variable selection via convex optimization. Ann. Appl. Stat. **9**(3), 1103 (2015)
Cachon, G.P., Swinney, R.: The value of fast fashion: quick response, enhanced design, and strategic consumer behavior. Manag. Sci. **57**(4), 778–795 (2011)
Cassel, E., Mendelsohn, R.: The choice of functional forms for hedonic price equations: comment. J. Urban Econ. **18**(2), 135–142 (1985)
Cavallo, A.: Are online and offline prices similar? evidence from large multi-channel retailers. Am. Econ. Rev. **107**(1), 283–303 (2017)
Cavallo, A.: Scraped data and sticky prices. Rev. Econ. Stat. **100**(1), 105–119 (2018)
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
Einav, L., Levin, J.: Economics in the age of big data. Science **346**(6210), 1243089 (2014)
Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. National Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
Nowak, A., Smith, P.: Textual analysis in real estate. J. Appl. Economet. **32**(4), 896–918 (2017)
Park, M.Y., Hastie, T., Tibshirani, R.: Averaged gene expressions for regression. Biostatistics **8**(2), 212–227 (2007)
Steyvers, M., Griffiths, T.: Probabilistic topic models. Handb. Latent Semant. Anal. **427**(7), 424–440 (2007)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal Stat. Soc.: Ser. B **58**(1), 267–288 (1996)
Tibshirani, R.J., Taylor, J., et al.: The solution path of the generalized lasso. Annal. Stat. **39**(3), 1335–1371 (2011)
Wainwright, M.J., Jordan, M.I., et al.: Graphical models, exponential families, and variational inference. Found. Trends® Mach. Learn. **1**(1–2), 1–305 (2008)
Yan, X., Bien, J.: Rare feature selection in high dimensions. J. Am. Stat. Assoc, pp 1–14, (2020)