JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. (2016)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2550

TEXTUAL ANALYSIS IN REAL ESTATE

ADAM NOWAK[a]* AND PATRICK SMITH[b]
[a] *College of Business and Economics, West Virginia University, Morgantown, WV, USA*
[b] *Department of Finance, San Diego State University, CA, USA*

SUMMARY

This paper incorporates text data from MLS listings into a hedonic pricing model. We show that the comments section of the MLS, which is populated by real estate agents who arguably have the most local market knowledge and know what homebuyers value, provides information that improves the performance of both in-sample and out-of-sample pricing estimates. Text is found to decrease pricing error by more than 25%. Information from text is incorporated into a linear model using a tokenization approach. By doing so, the implicit prices for various words and phrases are estimated. The estimation focuses on simultaneous variable selection and estimation for linear models in the presence of a large number of variables using a penalized regression. The LASSO procedure and variants are shown to outperform least-squares in out-of-sample testing. Copyright © 2016 John Wiley & Sons, Ltd.

Received 13 October 2015; Revised 9 May 2016


Supporting information may be found in the online version of this article.

1. INTRODUCTION

Real estate is one of the most studied asset classes—and for good reason. Some of the more prominent features of real estate include its incredible market value, the large share of real estate in individual investors' portfolios, and the value of mortgages tied to real estate. Even when focusing solely on households, the numbers are staggering. In 2014, household real estate assets were valued at $23.5 trillion USD, making up 24.2% of total household assets. Home mortgages on the household balance sheet were $9.4 trillion or 66.2% of total household liabilities.[1] For these reasons alone, researchers, policymakers, investors, homeowners and bankers all have a significant interest in accurately valuing real estate.

Valuation models for real estate can be derived using comparable sales, repeat sales, discounted cash flows or other means. This study uses a hedonic model where the price of a property is expressed as a linear combination of its attributes.[2] We argue that useful valuation models produce coefficients that are easily interpreted and provide pricing accuracy. The contributions of this paper are both methodological and empirical. The methodology described in this paper (i) applies textual analysis methods to real estate listings using a token approach and (ii) describes an estimation procedure that yields interpretable results. Empirically, the study finds (i) listing descriptions provided by listing agents contain information that can be used to decrease pricing error when used in conjunction with standard housing attributes in a hedonic pricing model, (ii) penalized regression outperforms a least-squares alternative in out-of-sample testing, and (iii) theoretically based penalty functions have similar performance to cross-validated penalized functions in out-of-sample testing.

* Correspondence to: Adam Nowak, College of Business and Economics, West Virginia University, Morgantown, WV, USA.
E-mail: adam.d.nowak@gmail.com

[1] http://www.federalreserve.gov/releases/z1/Current/z1.pdf, Table B.101.

[2] Early uses of the hedonic model include Rosen (1974). Two helpful literature reviews include Malpezzi (2003) and Kang and Reichert (1991).

The estimation technique described in this paper combines two branches of statistical research: textual analysis and sparse modeling. Textual analysis is a term for techniques that map text—news articles, 10-k's, message board postings, litigation releases, etc.—into variables that can be used in statistical applications including hypothesis testing, prediction and filtering. This study uses an approach whereby each remark can be expressed as a collection of words or phrases; each word or phrase is defined as a *token*. Tokens in the listing remarks can proxy for actual features of the property, seller or neighborhood. We are interested in selecting which tokens are relevant as well as the implicit prices for the features that they represent. In order to do so, indicator variables for tokens are included along with standard attribute variables in a linear, hedonic model. Because the number of tokens can increase with the total number of observations, the number of indicator variables for the tokens can be large. In such high-dimensional settings, least-squares estimation is at worst infeasible and at best prone to overfit the data, producing poor out-of-sample performance. Thus estimating the parameters requires techniques designed for large-dimensional parameter spaces.

One approach to high-dimensional data is to transform the data using data reduction methods. Data reduction techniques implicitly or explicitly assume that a large number of variables can be expressed using a much smaller set of observed or unobserved variables. One popular method for dimension reduction in linear models is principal components analysis (PCA). PCA creates principal components using linear combinations of a much larger set of variables from a multivariate dataset. Interpreting the coefficients on the principal components requires the researcher to first interpret the principal components, which can prove a challenge as all variables have non-zero loadings.

In a textual analysis setting where the data consist of token counts, topic modeling can be used to reduce the dimension of the data (Blei *et al.*, 2003). In these models, tokens in the document are assumed to come from one or more topics; alternatively, the document discusses one or more topics. For example, this paper discusses three topics: textual analysis, sparse modeling and real estate. Therefore, words or pharases relating to these topics are more likely to occur in the text than words specific to macroeconomics or international trade.

An alternative approach to dimension reduction is to assume that the true model is well approximated by a subset of explanatory variables. In the context of a linear regression, this implies the coefficient vector has some elements equal to 0. In this situation, the coefficient vector is said to be *sparse*. Estimating which coefficients are non-zero is variable selection. Traditional approaches such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) require estimating all combinations of models. Given the large number of tokens and the combinatorial nature of this approach, these approaches are computationally prohibitive.

The least absolute shrinkage and selection operator (LASSO) described in Tibshirani (1996) provides a feasible alternative. LASSO simultaneously performs model selection and coefficient estimation. Owing to a penalty function, coefficients are biased towards 0 but can still be consistent. Given the large number of observations in the dataset, biased but consistent coefficients can improve out-of-sample performance. In short, LASSO (i) screens for important tokens, (ii) provides easily interpreted coefficients, and (iii) performs well in out-of-sample testing—three features that are important when valuing real estate.

The remainder of the paper is organized as follows. Section 2 provides a literature review that emphasizes both sparse modeling and textual analysis. Section 3 describes some relevant theoretical results, the statistical techniques used, the details of the data source and the results from the estimation. Section 4 provides a summary of the paper and outlines areas for further research.

2. LITERATURE REVIEW

This study models residential property prices using a hedonic model. An important feature of the hedonic model is that property attributes explicitly impact property prices. Quantitative, qualitative,

DOI: 10.1002/jae

geographic and municipal attributes have all been found to influence property prices. Brown and Pollakowski (1977), Bond *et al.* (2002) and Rouwendal *et al.* (2016) find that water access and coastline significantly influence property prices. Benson *et al.* (1998), Paterson and Boyle (2002), Song and Knaap (2003) and Tu and Eppli (1999) find that non-traditional attributes can play a significant role. The running theme in all of these studies is that property price predictions can be improved by augmenting a simple hedonic pricing model (one that includes bedrooms, bathrooms, square footage, etc.) with non-standard attributes. Unfortunately, these non-standard attributes can be difficult or impossible for the researcher to measure. However, it is quite possible that these non-standard attributes are explicitly mentioned by listing agents in the remarks section of the listing. Hill *et al.* (1997) was one of the first studies to explicitly acknowledge that the remarks section in MLS data may contain 'hidden characteristics'. When constructing their repeat sales model Hill *et al.* (1997) use the remarks section to ensure house characteristics remained the same between sales.

Despite the frequent use of MLS data in real estate research, there are a very few studies that examine and include the non-standard attributes available in the MLS remarks section. In the previous studies that utilize the MLS remarks section, researchers have manually created indicator variables. Not only is this a time-consuming process, but also it is prone to human error. Haag *et al.* (2000) were the first to include the non-standard attributes available in the MLS remarks section in a hedonic model. They identify a list of keywords and phrases that were prevalent in their dataset (1994–1997) to examine the motivation of the seller, location of the property, physical improvements or property defects. In a recent follow-up study, Goodwin *et al.* (2014) extend the Haag *et al.* (2000) study by including additional keywords and categories. Goodwin *et al.* (2014) also cover a longer time period (2000–2009) that includes both an up and down real estate market. This is important because a study by Soyeh *et al.* (2014), which also utilizes the MLS remarks section, finds that incentives offered by sellers are not capitalized into sales price during soft market conditions.

Two approaches have been used when scoring or sorting text for use in financial and economics applications. The first approach pre-specifies positive and negative dictionaries of tokens and scores the text based on the relative frequencies of positive and negative tokens. Tetlock (2007), Loughran and McDonald (2011) and Bollen *et al.* (2011) find that text from the *Wall Street Journal*, 10-k filings and Twitter are all associated with future stock price movements. Garcia (2013) finds that the predictive power of text increases in recessions. In one of the few real estate applications, Pryce and Oates (2008) use a pre-specified dictionary approach and find real estate agents alter their choice of words based on the macroeconomy. It is important to emphasize that the dictionary approach is suitable only when the researcher has ex ante knowledge of relevant tokens for the application at hand. Loughran and McDonald (2011) emphasize this and show that a customized financial dictionary outperforms the general Harvard-IV-4 TagNeg dictionary when extracting information from 10-k's.

The second approach is a variant of supervised learning where a scored text is used to determine which tokens are more likely to increase a text's score. Using the US Congressional record, Gentzkow and Shapiro (2010) find that tokens can be used to identify Republicans and Democrats. Taddy (2013a) performs a similar analysis using information from Twitter. Taddy (2013b) rigorously studies a token model in a sparse coefficient setting. Mitra and Gilbert (2014) also seeks a sparse coefficient solution when searching for tokens that are associated with successfully funded projects on the crowdfunding website Kickstarter. We follow most closely the last three studies and use the sale price of the property as a way to identify a sparse coefficient vector.

Given the large number of potential tokens that can appear, we require a procedure for selecting which tokens are important that will not overfit the data. Common approaches for variable selection suggest comparing AIC, BIC or adjusted R^2 across models. To use any of these methods methods requires calculating 2^K models, where K is the number of candidate variables. Sala-i Martin (1997) examines the variable selection problem in the context of a cross-county economic growth equation and finds more than 3.4 billion models would need to be estimated when using 62 variables commonly

DOI: 10.1002/jae

used in the growth literature. Noting this computational burden, Fernandez *et al.* (2001) suggest a feasible Bayesian approach that provides a distribution over the more likely candidate models.

PCA is commonly used to reduce dimension when the likelihood function is normal. As mentioned in the Introduction, topic modeling is a more appropriate method for dimension reduction when the data, token counts in MLS listings, have a multinomial likelihood (Hofmann, 1999). Blei *et al.* (2003) describe latent Dirichlet allocation (LDA) where one or more topics are assigned weights in a given text and tokens are then drawn according to the distribution of tokens conditional on the assigned topics and weights. For these reasons, topic modeling is used as a means to classify text. The validity of this approach has been confirmed using scientific abstracts (Griffiths and Steyvers, 2004) and human subjects (Chang *et al.*, 2009). For a summary of other evaluation methods, see Wallach *et al.* (2009).

Sparse modeling has been used in engineering, statistics and economics applications; theoretical results draw from all disciplines. The LASSO (Tibshirani, 1996) estimates a linear model subject to a parametrized penalty function for the l_1 norm of the coefficient vector. Further modifications and discussions of the penalty functions can be found in Fan and Li (2001), Zou and Hastie (2005) and Belloni *et al.* (2011). The l_1 penalty sets less important coefficients to 0, thereby selecting a subset of variables for prediction. Knight and Fu (2000) provide asymptotic results for the LASSO for a fixed number of variables. Other authors provide asymptotic results when the number of variables is allowed to grow with the number of observations. Results include bounds and rates of convergence for prediction error (Greenshtein and Ritov, 2004; Bunea *et al.*, 2007; Bickel *et al.*, 2009), parameter estimates (Bickel *et al.*, 2009) and variable selection (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009).

An important choice for the researcher when using the LASSO is the specification and estimation of the penalty. A common approach is to use M-fold cross-validation (Varian, 2014). This technique is easy to perform, has been shown to perform well in Monte Carlo studies, but has also been shown to select too many irrelevant variables (Leng *et al.*, 2006; Zou *et al.*, 2007; Feng and Yu, 2013). Alternatively, the penalty can be based on theoretical results for prediction error or the coefficient vector (Knight and Fu, 2000; Bickel *et al.*, 2009). Belloni and Chernozhukov (2013) provide a feasible procedure for such an estimation. When the errors are no longer homoscedastic, Belloni *et al.* (2012) provide the required modification of the penalty function. Yet another approach is to use a square root loss function, described in Belloni *et al.* (2011), which results in a parameter-free penalty function. In order to more directly compare our results to the least-square procedure, we keep a squared loss function and use both the M-fold cross-validation procedure and procedures in Belloni *et al.* (2012) and Belloni and Chernozhukov (2013).

The use of LASSO and other penalized procedures becomes a valuable tool when the researcher is faced with a large number of regressors and would like to estimate a regression model. In such situations, the researcher must choose which variables to use based on a behavioral model, anecdotal evidence or other results in the literature. The methods discussed in this paper are applicable to these and other applications in finance and economics, when the researcher must select relevant variables in a linear model without any such guidance.

3. MODELING AND ECONOMETRIC ANALYSIS

3.1. Penalized Regression Theory

The data is an unbalanced panel. There are $i = 1, \ldots, I$ houses sold over time periods $t = 1, \ldots, T$, with some houses selling more than once for a total of $n = 1, \ldots, N$ transactions. For each transaction n, the sale price is given by

$$p_n = x_n \delta + \epsilon_n \tag{1}$$

DOI: 10.1002/jae

where p_n is the log of sale price, and x_n is a $1 \times K$ vector of attribute variables and indicator variables for the tokens.[3] δ is a $1 \times K$ vector of implicit prices for the variables and ϵ_{it} is an i.i.d. $N(0, \sigma^2)$ random variable capturing any variation in house prices not captured by the variables. Control variables include the square footage of living space, the lot square footage, the number of bedrooms, the number of bathrooms, the age of the property and indicator variables for foreclosure sales, real-estate-owned sales, short sales and agent-owned properties. Details for constructing the indicator variables for the tokens are described below. The linearity assumption in equation (1) is made for simplicity but not required.

When δ is sparse, $Q < K$ coefficients indexed by $S \subseteq \{1, \ldots, K\}$ are non-zero, and the remaining $K - Q$ coefficients are equal to 0. An alternative interpretation is that the true regression function is well approximated by a sparse δ, in which case ϵ_{it} includes an approximation error. All of the procedures described below are still viable with this alternative interpretation. The large number of variables in the model preclude using AIC or BIC methods in order to determine S, as there are more than 2^K possible models. However, Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) provide straightforward procedures for variable selection using p-values from a least-squares regression. Both procedures use an ex ante false-discovery rate, where the expected fraction of incorrectly selected variables is equal to ρ. Given the possibly large number of relevant tokens in the model, even conservative values of ρ will result in valuation models that include many falsely discovered regressors. As a feasible alternative to estimating a sparse δ, we solve the following optimization problem after standardizing all of the variables so that $\frac{1}{N}\sum_n x_{nk}^2 = 1$, where x_{nk} is the value of variable k for transaction n:

$$\min_d \frac{1}{N}\sum_{i,t}(p_{it} - x_{it}d)^2 + \frac{\lambda}{N}\sum_{k=1}^{K}|d_k| \tag{2}$$

Equation (2) is the sum of squared errors plus a penalty function for all coefficients excluding the intercept. The penalty function is proportional to the sum of the absolute values of the elements of d. The parameter λ is a tuning parameter, or weight, for the penalty function that controls the penalty for adding coefficients. Define $\hat{d}$ as the vector that minimizes equation (2), $\hat{Q}$ as the number of non-zero coefficients in $\hat{d}$, and $\hat{S}$ as the index of these non-zero coefficients. When $\lambda = 0$, the objective function in equation (2) is the least-squares objective function, and the minimizer is the least-squares estimator. The least-squares estimator does not provide a sparse solution as almost surely all entries in $\hat{d}$ are non-zero and $\hat{Q} = K$. When $\lambda > 0$, the estimator is the LASSO in Lagrangian form. Because of the shape of the penalty function, the LASSO estimator possesses a variable selection property in that it can provide a $\hat{d}$ with $\hat{Q} < K$.

Equation (2) cannot be solved using first-order conditions as the penalty function $\lambda\sum_{k=1}^{K}|d_k|$ is non-differentiable at $d = 0$. However, the problem can be recast into a Kuhn–Tucker maximization problem ensuring the solution is unique when $K < N$. Due to the penalty function in equation (2), the estimate $\hat{d}$ is biased towards 0 when $0 < \lambda$. When $\lambda = 0$ the resulting least-squares coefficient estimates are unbiased. However, due to the large number of variables, the variance of the least-squares coefficients can be large. LASSO makes a bias–variance trade-off in order to decrease out-of-sample prediction error. Alternative penalty functions for the coefficients are possible but are beyond the scope of this paper.[4]

Bühlmann and Van De Geer (2011) provide an overview of theoretical results and conditions for the LASSO relating to prediction error and variable selection when both K and N are large. Although we

[3] An intercept is included in the estimation but is omitted from the text in order to facilitate notation.

[4] When the penalty function uses the sum of squared coefficients instead of the sum of the absolute value of the coefficients, the resulting estimator is a ridge regression. The elastic net described in Zou and Hastie (2005) uses a penalty function that is a weighted combination of both the sum of the absolute value of the coefficients and the sum of the squared coefficients. Fan and Li (2001) use a quadratic spline penalty function.

DOI: 10.1002/jae

can always find a solution to equation (2), we would like to find conditions for which the prediction error of the resulting valuation model has desirable properties. For suitable λ and growth in $\|\delta\|_1$, predicted prices using $\hat{d}$ are consistent estimators of the true predicted price, $x_n\delta$. With additional assumptions on the explanatory variables and Q, an oracle equality can be shown where, up to a $\log(K)$ factor, the prediction error for the LASSO is comparable to that of least squares with ex ante knowledge of the true S. Thus the LASSO is well suited to the setting at hand where the valuation models include a large number of tokens.

In addition to estimating the price of a property, we would also like to identify which tokens are relevant when pricing real estate. Ideally, we would like to claim that the LASSO performs variable selection and correctly identifies S and signs δ with high probability as $N \to \infty$. However, the variable selection property requires non-trivial assumptions on the regressors, Q and $\min_{k\in S}|\delta_k|$. Required conditions for such a result can be found in Meinshausen and Bühlmann (2006), Zhao and Yu (2006) and Wainwright (2009). The intuition for the results is that when the variables in S are not too correlated with each other (minimum eigenvalue) or the variables not in S (irrepresentability), Q does not grow too quickly (sparsity), and $\min_{k\in S}|\delta_k|$ does not decay to 0 too quickly (delta-min), the true support is recovered in the limit with a high probability.[5]

A less ambitious goal—compared to variable selection—is variable screening, where the primary goal is to select all of the *substantial* variables with an absolute coefficient bounded away from 0. The substantial variables are a subset of the non-zero coefficients $S^* \subseteq S$ that have coefficients satisfying $C < |\delta_k|$. Identifying S^* is possible when $\hat{d}$ converges in probability to δ; alternatively, $S^* \subseteq \hat{S}$ with high probability in large samples. Convergence in probability requires suitable λ and Q as well as conditions on the regressors.[6] Although our dataset has a large number of observations, it is still a finite sample. With this in mind, we acknowledge that $\hat{S}$ can have a high false discovery rate and omit relevant variables with small impacts on price but still include substantial variables with a high probability. With this in mind, we emphasize that our results are exploratory in nature, and we do not claim to have selected all of the tokens relevant for real estate pricing.

3.2. Penalized Regression Procedures

Several techniques can be used both to select variables and to estimate coefficients. The Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) procedures select variables based on p-values from least-squares estimates. The LASSO and all of its variants minimize penalized least squares, and we collectively refer to the LASSO and all of its variants as *penalized estimators*. Not surprisingly, the value of the penalty plays a central role. An optimal choice of λ balances regularization and bias in $\hat{\delta}$. We investigate several choices of λ found in the literature. Cross-validated LASSO chooses a penalty in order to maximize out-of-sample root mean square error (RMSE). Belloni and Chernozhukov (2013) provide a data-dependent procedure for estimating the penalty and reducing bias in the coefficients. The procedure in Belloni *et al.* (2012) provides a procedure when there is heterscedasticity in the errors.

A common procedure for the choice of λ is to use an M-fold cross-validation (CV) procedure. The M-fold CV procedure uses subsets of the observations in order to estimate $\hat{d}$ and uses the remaining observations to determine out-of-sample performance. This process is repeated M times and the out-of-sample performance is averaged over all M trials. The CV penalty, λ_{CV}, is the value that provides the best out-of-sample performance and is calculated using the following procedure:

[5] Noting the restrictive nature of these conditions, Bühlmann and Van De Geer (2011) state that exact variable selection is only applicable for a 'rather narrow range of problems'.

[6] Necessary conditions for consistency are given in Knight and Fu (2000) for fixed K and in Bickel *et al.* (2009) for large K. Note that convergence of $\hat{d}$ to δ does not imply correct variable selection, as $\hat{d}$ might equal 0 for those $k \in S$ with $|\delta_k| \approx 0$.

DOI: 10.1002/jae

1. Sort all the observations into groups $m = 1, \ldots, M$ groups with an equal number of observations in each group.
2. For a given m and λ, estimate equation (2) using all observations not in subset m. Store the coefficients as $\widehat{d(\lambda)}^m$.
3. Estimate sale prices for sales in subset m using $\widehat{d(\lambda)}^m$ and store the total sum of squared errors (SSE) for group m as $\text{SSE}(\lambda)^m = \sum \left(p_i - z_i \widehat{d(\lambda)}^m \right)^2$.
4. The average SSE for this choice of λ is then $\text{SSE}(\lambda) = \frac{1}{M} \sum \text{SSE}(\lambda)^m$.
5. The value of λ that minimizes $\text{SSE}(\lambda)$ is the M-fold cross-validated λ_{CV}.

Empirically, λ_{CV} is often too small and too many variables are selected. This should not be surprising as CV aims to provide the best out-of-sample predictor for $\hat{d}$ and does not take into account the resulting properties of $\hat{d}$.

An alternative to CV is to use a λ that yields desirable theoretical results for $\hat{d}$. Bickel *et al.* (2009) show, for known σ and suitable regressors, that the penalty value $\lambda = 2\sigma\sqrt{2N \ln(KN)}$ results in $\hat{d}$ that converges in probability to δ. When σ is not known, Belloni and Chernozhukov (2013) provide a data-dependent procedure for a feasible estimation of the optimal λ. The idea is that the choice of λ must dominate the noise in equation (2) given by the gradient of the mean sum of squared errors evaluated at the true δ, $G = 2E_N[x_n \epsilon_n]$. In order to do so, the researcher chooses $c > 1$, $\alpha = o(N^{-1})$, a small, positive η (or a maximum number of steps J), and an initial condition $\hat{\sigma}^{j=0}$ equal to the sample standard deviation of p_n. Next, the parameter-free, data-dependent value $\|G/2\sigma\|_\infty$ is simulated using the relationship

$$\|G/2\sigma\|_\infty =_d \max_{1 \le k \le K} |E_N[x_{nk} g_n]|, g_n \ N(0,1) \tag{3}$$

From these simulations, the $1 - \alpha$ sample quantile of $N\|G/2\sigma\|_\infty$ is computed, $\Lambda(1 - \alpha|X)$. This quantile is then used to estimate λ using the following procedure. Using superscripts to indicate a particular iteration $j = 1, \ldots, J$:

1. Set $\lambda^j = 2c\hat{\sigma}^{j-1}\Lambda(1 - \alpha|X)$.
2. Estimate equation (2) using λ^j and store $\hat{d}^j$ and $\hat{Q}^j$.
3. Update $\hat{\sigma}^j$ as $\sqrt{\frac{N}{N - \hat{Q}^j} \sum_i \left(p_i - x_i \hat{d}^j \right)^2}$.
4. If $\left| \hat{\sigma}^j - \hat{\sigma}^{j-1} \right| < \eta$ or $j = J$, stop.
5. Otherwise, increase j to $j + 1$ and repeat steps 1–4.

Following Monte Carlo simulations given in Belloni and Chernozhukov (2013), we set $c = 1.1$ and $\alpha = 0.1$. The above procedure results in a consistent estimate of σ that can also be used to create a feasible optimal penalty $\lambda_F = 2c\hat{\sigma}\sqrt{2N \ln(2K/\alpha)}$. However, Belloni and Chernozhukov (2013) recommend the data-dependent penalty level $\lambda_{\text{DD}} = 2c\hat{\sigma}^{s-1}\Lambda(1 - \alpha|X)$ be used as it adapts to the regressors and is less conservative, in that it has a smaller penalty $\lambda_{\text{DD}} \le \lambda_F$. Either penalty level can be used to estimate $\hat{d}$.

As mentioned above, the LASSO procedure results in biased $\hat{d}$. The post-LASSO procedure described in Belloni and Chernozhukov (2013) is a two-stage procedure that mitigates this bias. In the first stage, LASSO selects $\hat{S}$. In the second stage, the variables in $\hat{S}$ are used in a least-squares regression. The post-LASSO procedure is outlined in the following three steps:

1. For a given λ, estimate equation (2).
2. Create a $1 \times \hat{Q}$ vector x_n^{PL} by removing the $K - \hat{Q}$ variables in x_n that are not in $\hat{S}$.
3. Set $\lambda = 0$ and estimate equation (2) using only the variables in x_n^{PL}.

DOI: 10.1002/jae

By setting $\lambda = 0$, the second-stage estimation procedure becomes least squares. In the first stage, the value of λ can be estimated using the iterative procedure in Belloni and Chernozhukov (2013). The intuition for this procedure is the following: if LASSO correctly estimates S in the first stage, the model in the second stage is correctly specified, and the resulting least-squares coefficients are unbiased. For reasons mentioned above, it is likely that $S \neq \hat{S}$, and the second-stage model is misspecified. Despite this, Belloni and Chernozhukov (2013) find that erroneously included or excluded variables will have little explanatory power.

The previous procedures assume homoscedastic errors. When errors are homoscedastic, a modification of the standard LASSO given in Belloni *et al.* (2012) can be used. The procedure permits non-Gaussian and heteroscedastic errors with the additional assumption $\log K = o(N^{1/3})$. The procedure modifies the objective function in equation (2):

$$\min_d \frac{1}{N}\sum_{i,t}(p_{it} - x_{it}d)^2 + \frac{\lambda}{N}\sum_{k=1}^{K}|\psi_k d_k| \tag{4}$$

Here, the coefficients in d have unique penalty weights given by ψ_k; however, the shape of the penalty is unchanged and the solution is still sparse. The weights are estimated using a data-driven algorithm. For iteration $j = 1, \ldots, J$, $c > 1$, and η a small positive number, the ψ_k are estimated using the following algorithm:

1. Calculate residuals $\hat{e}_n^j = p_n - x_n\hat{d}^{j-1}$, if $j = 1$ set $\hat{e}_n^j = p_n - \frac{1}{N}\sum_n p_n$.
2. Calculate penalty terms $\hat{\psi}_k^j = \sqrt{\frac{1}{N}\sum_i x_{kt}^2 \hat{e}_n^{j2}}$.
3. Solve equation (4) using $\hat{\psi}_k^j$ and $\lambda = 2c\sqrt{N}\Phi^{-1}\left(1 - \frac{\alpha}{2K}\right)$.[7]
4. If $\left\|\hat{d}^j - \hat{d}^{j-1}\right\| < \eta$ or $j = J$, stop.
5. Otherwise, increase j to $j + 1$ and repeat steps 1–4.

The above procedure results in an estimate of $\hat{d}$ that is valid in the presence of non-Gaussian and heteroscedastic disturbances, a common occurrence in economics data sets. Other scalers besides $\lambda = 2c\sqrt{N}\Phi^{-1}\left(1 - \frac{\alpha}{2K}\right)$ are possible. Required conditions for alternative λ are given in Belloni *et al.* (2012).

3.3. Alternative Pricing Models

We use five models as a means to compare the relative and supplemental predictive power contained in the remarks. In this section, we use subscripts i, t emphasizing the panel structure of the data:

$$[\text{BASE}] : p_{it} = \alpha_t + \gamma_z + h_{it}\beta + \epsilon_i \tag{5}$$

$$[\text{U1}] : p_{it} = \alpha_t + h_{it}\beta + v_i\theta + \epsilon_i \tag{6}$$

$$[\text{B1}] : p_{it} = \alpha_t + h_{it}\beta + w_i\phi + \epsilon_i \tag{7}$$

$$[\text{U2}] : p_{it} = \alpha_t + \gamma_z + h_{it}\beta + v_i\theta + \epsilon_i \tag{8}$$

$$[\text{B2}] : p_{it} = \alpha_t + \gamma_z + h_{it}\beta + w_i\phi + \epsilon_i \tag{9}$$

[7] Here, $\Phi^{-1}(z)$ is the inverse cumulative distribution function for the standard normal distribution.

DOI: 10.1002/jae

Here, α_t is a time fixed-effect for time period t, h_{it} is a vector of control variables that includes the number of bedrooms, bathrooms, square footage of living space, lot square footage and indicator variables for the nature of the sale, v_i is a vector of indicator of variables for the unigrams and w_i is a vector of indicator variables for the bigrams. Construction of the unigram and bigram vectors is described in the following section. The vector β contains relative prices for the control variables. γ_z is a census tract fixed effect for all 598 census tracts with 10 or more sales. The vectors θ and ϕ are implicit prices of the tokens. Finally, ϵ_i is an i.i.d., normally distributed error term $N(0, \sigma^2)$. It is fully acknowledged that ϵ_i includes the effect of any unobserved variable not mentioned in the remarks that can be related to the property attributes or the nature of the transaction as well as any approximation error ascribed to the functional form.

Equation (5) is the baseline model that includes the control variables as well as time and location fixed effects. Such a model is commonly used in the real estate literature. Census tract fixed effects are used in order both to control for unobserved variation in quality due to location and remove the effect of any explicit location effects mentioned in the remarks. After experimenting with several configurations for the control variables, we found that equation (5) produced R^2 values that were comparable to R^2 values from other model specifications of comparable complexity. We make no claim as to the unbiasedness of the estimates for β but note that the explanatory power of the specification in equation (5) is comparable to the explanatory power of alternative models using transformations and interactions of the control variables.

Equations (6) and (7) regress price on control variables and tokens in the absence of census tract fixed effects. These models are used to assess whether information in remarks can substitute for the information conveyed in census tract fixed effects. In these models, relevant tokens are picking up information related to location. Equations (8) and (9) are constructed in order to highlight the supplemental information tokens can provide. Assuming census tract fixed effects capture all information relevant to location, relevant tokens are now picking up information that is property specific. More elaborate interactions between tokens and control variables are possible but are beyond the focus of this paper.

We also apply our approach to another popular pricing estimator in the real estate literature. The *repeat sales regression* regresses differenced sale prices on differenced right-hand-side variables. For consecutive sales of the same house i sold at times s and $s \leq t$, the change in price, $\Delta p_{it} = p_{it} - p_{is}$, is given differencing equation (1):

$$\Delta p_{it} = \Delta x_{it}\delta + \Delta\epsilon_{it} \tag{10}$$

Here, Δx_{it} is the difference in right-hand-side variables. When x_{it} contains only time period fixed effects, Δx_{it} contains 0's, a $+1$ for the time period t variable and a -1 for the time period s variable. The repeat sales regression treats time-invariant variables as nuisance parameters. Such time-invariant variables include location effects and possibly structural effects when quality does not change. Implicit in the repeat sales regression is the assumption that the quality of the underlying property does not change. With this assumption, the coefficients on the time effects are interpreted as a constant-quality price index. When tokens are included in the repeat sales regression, relevant tokens capture time-varying information related to a specific property as all time-invariant effects have been removed.

Remarks associated with two different sales of the same house are almost surely time-varying, although certain features of the underlying property are time invariant. When we include tokens in x_{it}, the effects of time-invariant tokens are differenced away. However, certain relevant features of the property are both time variant and indicated in the remarks. For example, renovating a property would presumably increase the sale price; properties that are recently renovated would have larger changes in

DOI: 10.1002/jae

Table I. Sample MLS listing

Zip code	Beds	Baths	Sale date (m/d/yy)	Sale price	Remarks
30043	3	2	6/7/13	$270,000	back on market!!! located in tranquil neighborhood with sought-after schools close to shopping and i-85. this 3 bedroom 2 bath home is beautifully decorated. new roof was installed 3/20/14. marble master bath is stunning. room for expansion upstairs
30043	3	2	6/16/13	$168,900	wonderful updated one level with vaulted great room w fireplace & gas logs, formal dining room, kitchen with corian, newer stove & microwave, breakfast area overlooks wooded backyard, master bedroom suite w/upgraded master bath with tiled shower & jetted tub
30043	3	2	6/17/13	$150,000	great new listing on 18th fairway of collins hill golf course on cul de sac too no hoa not a short sale and not bank owned pride of ownership here new double pane windows new roof updated heat and air gourmet kitchen with double gas oven ss fridge
30043	3	2	5/1/13	$113,500	adorable fannie mae homepath ranch style home updated and like new with new kitchen appliances, freshly painted, new carpet. large open living room with vaulted ceiling and fireplace, kitchen is spacious with breakfast area, nice master bathroom with tub shower
30043	3	2	6/16/13	$109,000	4 sided brick ranch with full basement. quick access to i85, 316, mall of ga. large family room w/fireplace, separate living room and dining room, kitchen w/eat in b'fast room, laundry room, two car carport, deck on back. huge fenced in backyard for kids.
30043	3	2	4/1/13	$96,000	cute 3 bed 2 bath 2-story home in cul-de-sac. great schools & great location. private fenced backyard. needs carpet & paint. short sale. hurry before it's gone. sold as is no repairs.
30043	3	2	4/1/13	$93,000	nice ranch-style home on level, wooded, fenced corner lot! vaulted, sun-filled great room with dining area with wood-laminate floors! master bedroom has full, private bath. single car carport & charming front porch. back yard has large walk-in shed. excellent.
30043	3	2	5/8/13	$86,125	3 bdr 2bth split level home that has tons of potential. great opportunity for investor or first time buyer willing to put in some sweat equity. great location close to shopping and sought after peachtree ridge high school.

prices than non-renovated properties. If macroeconomic factors lead to citywide renovation, the time coefficients are biased and no longer result in a constant-quality index.

The advantages of including tokens in the repeat sales regression are threefold. First, tokens can be used to mitigate bias in the price index by controlling for time-varying changes in quality. Second, prices of individual tokens can be used to estimate price differential based on listing agent assessments of quality. Third, when included alongside indicator variables for auctions, foreclosures or other events most likely associated with changes in quality, we can obtain unbiased coefficients in the presence of both time varying and time invariant. Mayer (1998) uses a repeat-sales approach to estimate auction premiums that control for unobserved time invariant. Because time-varying controls are not available, the auction premium in Mayer (1998) is presumably biased due to associated time-varying changes in quality associated with auction properties.

3.4. Tokens

Table I presents a sample of eight listings for three-bedroom two-bathroom houses in zip code 30043. The sale prices range from $270,000 to $86,125. Based on zip code, bathroom and bedroom it is impossible to explain variation in sale prices. However, the remarks for the property with the largest sale price indicate positive, unobserved features about the location (*located in tranquil neighborhood*) and the property itself (*marble master bath*). These remarks are in contrast to the property with the smallest sale price. There, the remarks indicate the property is not in great condition as the remarks indicate that the buyer must be *willing to put in some sweat equity*.

DOI: 10.1002/jae

The public remarks are processed in order to produce a set of variables that indicate certain tokens are present in the remarks. It is possible to create indicator variables for each word in the remarks. In the textual analysis literature, single words are called *unigrams*. Examples of unigrams include *ceiling* and *gated*. In addition to unigrams, this study also examines the use of *bigrams*. A bigram is a two-word phrase such as *drop ceiling*, *vaulted ceiling*, *gated windows* or *gated community*.

Before creating the bigrams, *stop words* are removed from the remarks section using a custom set. Stop words are words that are assumed not to convey any information about the property. A list of stop words specific to the remarks section, and real estate at large, is created. The list of stop words is available from the authors upon request. An additional step called *stemming* is often carried out in textual analysis. In unreported results, we found that generic stemming using the *SnowballC* package in R did not improve performance or change any of the results in the paper in a substantial manner. Therefore, for the sake of simplicity, the remarks were not stemmed. However, we found in the data that real-estate agents use various spellings and abbreviations for the same word. For instance, we find that *bedroom, bdrm, bdr, bd room, and bedrm* are all used. Therefore, we used a data-specific stemming program to map all such variations of this and other objects to *bedroom*. A complete list of such mappings is too voluminous to report but is available from the authors upon request.

The token approach models each remark as a collection of tokens. For all unigrams v_j, $j = 1, \ldots, J$, define the indicator variable $\mathbb{1}(v_j)_i = 1$ if unigram v_j is in remark i and 0 otherwise. The $1 \times J$ vector v_i is then defined as $v_i = (\mathbb{1}(v_1)_i, \ldots, \mathbb{1}(v_J)_i)$. A similar procedure is used to create the $1 \times J$ vector for bigrams, $w_i = (\mathbb{1}(w_1)_i, \ldots, \mathbb{1}(w_J)_i)$. Prices for the unigrams and bigrams are contained in the $1 \times J$ vectors $\theta = (\theta(v_1), \ldots, \theta(v_J))$ and $\phi = (\phi(v_1), \ldots, \phi(v_J))$, respectively.

Two alternatives to the above approach are also possible. The first approach uses counts and replaces the indicator function with the total number of times the token appears in remark i; the second approach uses frequencies rather than counts and replaces the indicator function with the total number of times the token appears in remark i, divided by the total number of tokens in remark i. In order to facilitate interpretation of the coefficients, we use the indicator function approach but note that in several experiments the results were robust to these two alternative approaches. In the context of equations (6)–(9), interpreting the coefficients in θ and ϕ is straightforward. Including u_j in the remarks increases (decreases) the expected price by an amount θ_j if $\theta_j > 0$ ($\theta_j < 0$).

If the researcher is not interested in the prices of tokens but rather aggregating the information contained in the remarks, the inner product $q_i = v_i\theta$ can be used. If we assume that the remarks contain information about house quality, we can interpret q_i as an index of quality. Furthermore, this index of quality can be used as a measure of quality in other regressions. A similar approach using a sufficient reduction paradigm is taken in Taddy (2013a, b).

However, it should be emphasized that the tokens are considered exchangeable in that the ordering of the tokens is not important for pricing purposes. For example, when using unigrams, the phrases *gated windows* and *gated community* will be priced as $\theta(gated) + \theta(windows)$ and $\theta(gated) + \theta(community)$, respectively. The difference in price between these two phrases is equal to $\theta(windows) - \theta(community)$. This is counter-intuitive as differences in housing quality indicated by *gated windows* and *gated community* come from the adjective *gated* modifying the nouns *windows* and *community*. Using bigrams alleviates issues associated with unigram exchangeability by capturing some notion of word ordering. When using bigrams as token, the difference in price between *gated windows* and *gated community* is equal to $\phi(gated\ windows) - \phi(gated\ community)$.

Without loss of generality, we use j to indicate the rank of the frequency of the token in the remarks. For example, $j = q$ is the most frequent token, $j = 2$ is the second most frequent token and so on. For practical purposes, it is necessary to truncate the list of total tokens available to use. First, it is not hard to rationalize that a token that appears in only one record is unlikely to appear in future records. It is also unlikely that this token can be used to predict future prices. Second, in order to compare the penalized procedures to least squares, we require the matrix of regressors to have full rank, which

DOI: 10.1002/jae

ensures least squares is feasible. We find the full rank condition is frequently violated when we choose $2000 < J$ using subperiods of the data. We experiment with several alternatives for J, including $J \in \{100, 500, 1000, 2000, 3000\}$.[8]

3.5. Data Description

The primary data source used in this study comes from the Georgia Multiple Listings Service (GAMLS). The GAMLS data include all single-family detached houses that were listed, regardless of whether they sold or not, from 1 January 2000 to 31 December 2014 in the five counties that form the core of the Atlanta metropolitan market (Fulton, Dekalb, Gwinnett, Cobb, and Clayton). In addition to the public remarks field described earlier, the GAMLS dataset includes information on the location and size of the property (address, acreage, etc.), physical characteristics of the house (number of bedrooms, bathrooms, etc.), and details of the transaction (listing date, listing price, sales price, etc.). All the data in the GAMLS is manually entered by a listing agent. Thus it is prone to error (Levitt and Syverson, 2008). It also does not include each house's square feet of living area. We circumvent these potential issues with data obtained from each county's tax assessor office. The tax assessor data include detailed parcel-level information that we use to determine the square feet of living area for each house in our study and validate the information in the GAMLS. The initial GAMLS dataset includes 511,851 listings. We apply several filters to systematically clean the data. First, we remove listings with missing or incomplete data. We then winsorize the top and bottom 0.5% of sales price to remove potential outliers. Finally, we exclude houses that were built before 1900, have less than 500 square feet of living area, or have 10 or more bedrooms. We apply these filters to limit the variability in our data and ensure it is reasonably homogeneous, as suggested by Butler (1980). The cleaned dataset includes 414,404 unique sales transactions. Descriptive statistics for the entire data are displayed in Table II.

4. RESULTS

4.1. Least-Squares and Sparse Estimation Comparison

Figures 1 and 2 display the positive and negative coefficients for the tokens with the largest magnitudes. The coefficients were estimated using the cross-validated LASSO procedure in equation (9) for the entire sample period (2000–2014), including census tract and time dummy variables. Coefficients with a larger magnitude are illustrated in larger font sizes. Overall, the terms in the figures suggest that the tokenization procedure can identify relevant phrases in the MLS remarks section that can be used in pricing models. A detailed list of the top 50 unigrams and bigrams sorted by magnitude is available in the online Appendix in Tables A2 and A3.

In the following tables, panel A presents the in-sample RMSE results, panel B presents the out-of-sample RMSE results and panel C presents the number of variables selected when calculating RMSE. The columns correspond to the RMSE when estimating models in equations (6)–(9) using ordinary least-squares (LS) (Benjamini and Yekutieli, 2001), false discovery rate (FDR), cross-validated LASSO (CV), Belloni and Chernozhukov feasible LASSO (BC), Post-LASSO (Post) and heteroscedastic LASSO (Het) procedures discussed above. In each model, a maximum of 2000 tokens are used. BASE includes a maximum of 598 census tract fixed effects and is always estimated using least squares. Panels A and B display the RMSE values for each model (equations (5)–(9)). When calculating the RMSE used in panel B, the $\hat{Q}$ variables in panel C are used. Only LS uses all $\hat{Q} = K$ variables. The other four procedures select the $\hat{Q} \leq K$ variables that are used to calculate

[8] Figure A1 in the online Appendix (supporting information) shows the cumulative distribution function for the 4000 most frequent tokens across all listings in the dataset. Counts for the less frequent tokens are quite large. The 4000th least frequent unigram (bigram) occurs 65 (220) times in the remarks. In our analysis, we use the 2000 most frequent tokens. The 2000th least frequent unigram (bigram) occurs 249 (453) times in the remarks.

DOI: 10.1002/jae

Table II. Descriptive statistics

	Min.	Mean	Median	Max.	SD
Panel A: 2000–2014					
Sale price ($1000s)	11.4	193	156.5	1099	143.8
List price ($1000s)	1	199	159.9	3400	151.1
Area (ft^2)	506	2220.1	2009	16475	979.4
# of bedrooms	1	3.6	4	9	0.9
# of bathrooms	1	2.3	2	12	0.8
Construction year	1900	1983.5	1989	2014	20.6
Sale year	2000	2006.9	2007	2014	4.1
Panel B: 2000–2007					
Sale price ($1000s)	11.5	202.5	165	1097.2	124.5
List price ($1000s)	13.9	207	168.2	2219	129.4
Area (ft^2)	520	2184.3	1994	16000	928.9
# of bedrooms	1	3.6	3	9	0.8
# of bathrooms	1	2.2	2	10	0.8
Construction year	1900	1983.6	1990	2007	20.1
Sale year	2000	2003.8	2004	2007	2.2
Panel C: 2008–2014					
Sale price ($1000s)	11.4	180.7	133	1099	164.9
List price ($1000s)	1	188.6	137.9	3400	174.9
Area (ft^2)	506	2266.9	2030	16475	1039.8
# of bedrooms	1	3.7	4	9	0.9
# of bathrooms	1	2.4	2	12	0.9
Construction year	1900	1983.5	1989	2014	21.2
Sale year	2008	2011	2011	2014	2
Panel D: 2012–2014					
Sale price ($1000s)	11.5	203.4	153	1099	172.9
List price ($1000s)	4	209.9	159	1790	181
Area (ft^2)	506	2316.3	2100	11525	1028.4
# of bedrooms	1	3.7	4	9	0.9
# of bathrooms	1	2.4	2	12	0.9
Construction year	1900	1983.2	1988	2014	21
Sale year	2012	2012.9	2013	2014	0.8

RMSE in panel B. By doing so, the reported RMSEs emphasize differences in RMSE due to bias and precision in the coefficient estimates alone and not differences in $\hat{Q}$. The number of observations in each period are listed in Table A.I.

We include results for several time frames. The first row in each panel includes data for the entire sample period (2000–2014). We then partition the data into pre-crash (2000–2007) and post-crash (2008–2014) subperiods to examine whether the in- and out-of-sample results are sensitive to the time period selected. Finally, in the last row of each panel we partition the data into a more recent subsample that includes results for 2012–2014. We include the smaller, more recent subsample for two reasons. First, while working with the MLS remarks data we noticed that a small percentage of remarks were truncated prior to 2012.[9] Second, we want to ensure that functional obsolescence does not impact the model results. Functional obsolescence in real estate occurs often as the desirability or usefulness of an attribute changes or becomes outdated. Thus a token's magnitude and sign may change over time if the attribute becomes functionally obsolete. Given the extended time period of our study, we include the 2012–2014 subsample to ensure functional obsolescence does not significantly impact the results.

The initial results provide an indication as to the information content of the tokens relative to location fixed effects. Table III compares the BASE model with census tract fixed effects to hedonic models

[9] We estimate that the data truncation affected less than 1% of the records prior to 2012. The small percentage of records that were affected were missing fewer than 16 characters each, which represents less than 7% of their total length. A discussion with our contact at GAMLS revealed that a systems upgrade was performed in the beginning of 2012 and was likely the source of the truncation.

DOI: 10.1002/jae

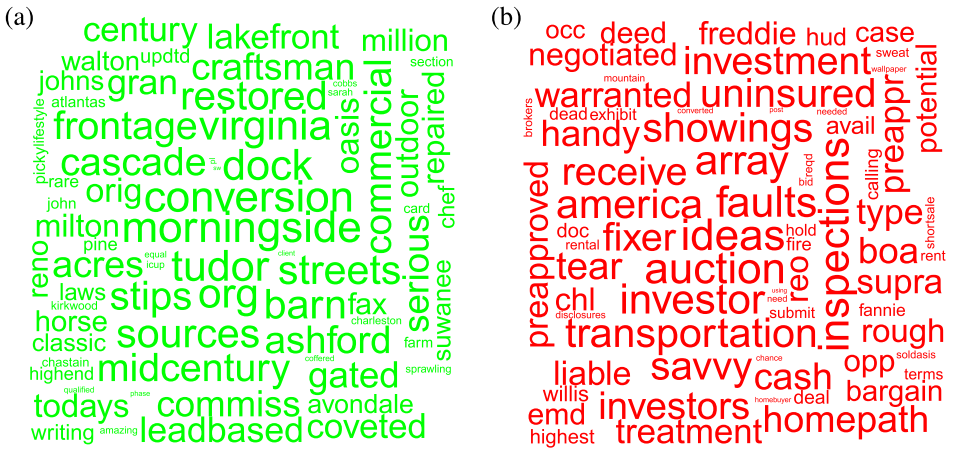

Figure 1. Unigrams: (a) positive Unigrams; (b) negative Unigrams

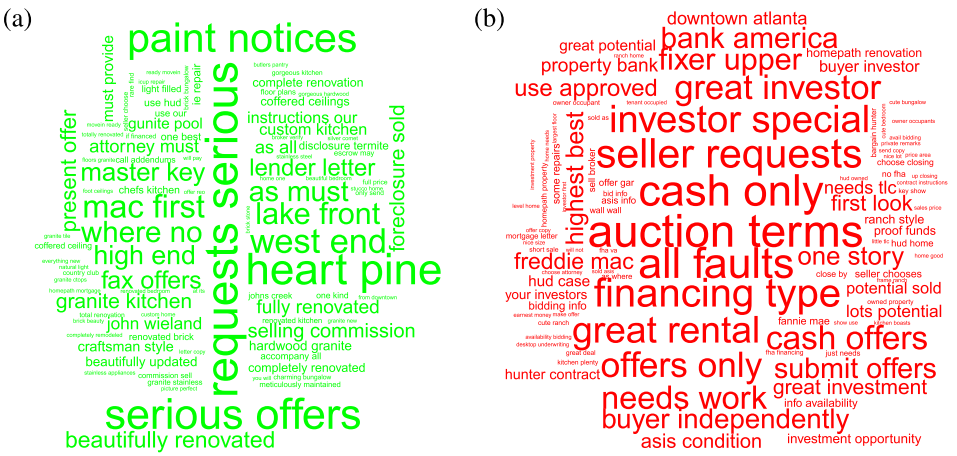

Figure 2. Bigrams: (a) positive bigrams; (b) negative bigrams

with tokens but without census tract fixed effects. Panel A indicates that the census tract fixed effects have more in-sample predictive power than the tokens in every time period. By definition, the LS model has the smallest in-sample RMSE of all models that include tokens. Moving across the columns in panel A, we find that the RMSE when using unigrams is always less than the RMSE when using bigrams. Further, the RMSE for U1 and B1 are comparable across the penalized procedures.[10]

Because least squares is guaranteed to minimize the in-sample RMSE, we are more interested in comparing out-of-sample RMSE. Owing to the large number of variables, we would expect LS to overfit the data in-sample, resulting in poor out-of-sample performance. We use a standard M-fold out-of-sample testing procedure in order to evaluate out-of-sample performance; details are provided in the online Appendix. The results in panel B indicate that the BASE model produces out-of-sample

[10] For computational feasibility, we use a grid search to calculate the CV penalty parameter. As a result, the CV does not always produce the smallest out-of-sample RMSE.

DOI: 10.1002/jae

Table III. Hedonic: log price without census tract fixed effects

		LS		FDR=0.1		CV		BC		Post		Het	
	(BASE)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)
Panel A: In-sample RMSE													
2000–2014	***0.329***	0.376	0.409	0.376	0.410	0.376	0.409	0.383	0.416	0.378	0.411	0.388	0.419
2000–2007	***0.199***	0.243	0.266	0.244	0.268	0.243	0.266	0.251	0.274	0.246	0.269	0.255	0.276
2008–2014	***0.356***	0.460	0.506	0.461	0.508	0.464	0.508	0.477	0.523	0.466	0.513	0.485	0.530
2012–2014	***0.318***	0.426	0.477	0.431	0.484	0.440	0.484	0.453	0.506	0.439	0.490	0.464	0.513
Panel B: Out-of-sample RMSE													
2000–2014	0.396	***0.378***	0.411	0.379	0.412	***0.378***	0.411	0.385	0.418	0.380	0.413	0.389	0.421
2000–2007	***0.202***	0.246	0.268	0.246	0.271	0.246	0.268	0.253	0.275	0.249	0.271	0.256	0.277
2008–2014	***0.410***	0.465	0.510	0.467	0.513	0.480	0.515	0.480	0.526	0.471	0.515	0.487	0.532
2012–2014	0.461	***0.437***	0.489	0.442	0.494	0.458	0.493	0.459	0.511	0.447	0.497	0.467	0.517
Panel C: $\hat{Q}$													
2000–2014		2024	2024	1436	1295	1923	1887	861	753	861	753	675	***598***
2000–2007		2016	2016	1272	1085	1837	1828	653	535	653	535	481	***454***
2008–2014		2016	2016	1196	1084	1267	1376	691	625	691	625	524	***487***
2012–2014		2012	2012	944	791	847	1104	539	427	539	427	385	***329***

Note: All models include annual fixed effects and control variables. A maximum of 2000 tokens are used. Minimum values for each subperiod are indicated with bold italics.

DOI: 10.1002/jae

RMSEs that are larger comparable to their in-sample RMSEs. The out-of-sample RMSEs increase dramatically for the entire period and for 2012–2014. By comparison, the models that use tokens instead of census tract fixed effects do not display significant changes in out-of-sample performance; this is true regardless of the estimation procedure. Because the number of census tracts is small compared to the number of tokens, we conclude that variation in house prices due to location is less precisely estimated than variation due to information that is captured by the tokens. However, the results in panel B indicate that information attributable to location and all information captured by the remarks are similar in magnitude.

Panel C displays the in-sample $\hat{Q}$ for each estimation procedure. Of particular note is the large drop in relevant tokens when using the BC and Het procedures. The BC and Het models produce in-sample and out-of-sample RMSEs that are comparable to the other estimation procedures that use more tokens. These results suggest that we can discard a large number of tokens when pricing real estate without sacrificing predictive power.

In Table IV we present the results of U2 and B2 and regress price on both tokens, control variables and census tract fixed effects. Comparing BASE to LS in panel A of Table IV, we find a decrease in in-sample RMSE between 2% and 5%. The in-sample RMSE of LS is comparable to the in-sample RMSE when using FDR or penalized procedures. However, results in panel B indicate that least-squares coefficients, either LS or FDR, have poor out-of-sample performance. Comparing LS in panels A and B, RMSE can increase significantly. In contrast, out-of-sample RMSEs for the penalized procedures are comparable to in-sample RMSEs. The Post estimator has performance comparable to CV in the first two subperiods and significantly outperforms all other estimators in the final two subperiods. The results in Table IV indicate that information in the remarks can provide valuable supplemental information when pricing real estate.

All of the hedonic models in equations (6)–(9) include both time-invariant and time-variant control variables and tokens. The results in Tables III and IV document the explanatory power of tokens associated with both time-invariant and time-varying attributes. In order to separately estimate the explanatory power of tokens associated with time-varying attributes, we difference equations U1 and B1 using same-property sales. This results in an augmented repeat-sales model. The repeat-sales model expresses changes in sale prices as changes in indicator variables associated with time. By differencing both sides of U1 and B1 and assuming constant implicit prices over time, we remove the effect of any time-invariant, unobserved variables including census tract fixed effects. We do not ex ante identify which tokens are associated with time-invariant variables but instead include all tokens when estimating the repeat sales regression. However, this does not present a problem as the coefficients on tokens associated with time-invariant attributes should be close to 0.

The results for the differenced U1 and B1 using same-property sales are displayed in Table V. The results in panel A of Table V show that tokens improve in-sample RMSE in every period. Panel B indicates that out-of-sample RMSE can also be improved when a sparse estimation procedure is employed. Panel C of Table V indicates that only a few tokens are selected by the BC, Post and Het procedures, but results in panel B show that these procedures do not show a significant difference in out-of-sample RMSE compared to the CV estimator. Comparing the results in Table V to Tables III and IV, we conclude that a large number of tokens capture time-invariant attributes; alternatively, it is possible to interpret this as mild evidence of the repeat sales estimator as a constant-quality house price index. In contrast, it is interesting to note the significant performance gain of the tokens in the subperiod 2008–2014. It is well documented that during this period housing prices were declining and distressed sales were increasing. To the extent that information in the remarks convey the distressed nature of the property, it appears that including tokens of this nature can be used to correct for any bias associated with the unobserved distressed nature of the property.

In the online Appendix, we provide results for how the tokens perform when N is small and estimate models U2 and B2 for each of the 15 years in our sample (Table A4). Three features similar to panel B

DOI: 10.1002/jae

Table IV. Hedonic: log price with census tract fixed effects

		LS		FDR=0.1		CV		BC		Post		Het	
	(BASE)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
Panel A: In-sample RMSE													
2000–2014	0.329	***0.286***	0.296	0.287	0.298	0.287	0.296	0.297	0.306	0.289	0.299	0.307	0.316
2000–2007	0.199	***0.173***	0.181	***0.173***	0.182	***0.173***	0.181	0.183	0.191	0.175	0.183	0.196	0.204
2008–2014	0.356	***0.304***	0.316	0.305	0.318	0.328	0.354	0.326	0.337	0.309	0.321	0.352	0.362
2012–2014	0.318	***0.268***	0.281	0.271	0.285	0.326	0.351	0.311	0.321	0.278	0.291	0.372	0.383
Panel B: Out-of-sample RMSE													
2000–2014	0.396	0.429	0.437	0.428	0.436	***0.288***	0.298	0.300	0.309	0.292	0.301	0.311	0.320
2000–2007	0.202	0.177	0.186	0.177	0.187	***0.175***	0.183	0.187	0.194	0.178	0.185	0.200	0.207
2008–2014	0.410	0.402	0.404	0.403	0.405	0.372	0.386	0.334	0.343	***0.314***	0.325	0.360	0.370
2012–2014	0.461	0.506	0.534	0.508	0.534	0.367	0.372	0.324	0.333	***0.286***	0.297	0.386	0.397
Panel C: $\hat{Q}$													
2000–2014		2622	2622	1886	1707	2473	2460	1217	1113	1217	1113	1052	***995***
2000–2007		2604	2604	1597	1444	2376	2288	987	915	987	915	851	***841***
2008–2014		2608	2608	1572	1500	1050	***853***	1064	1043	1064	1043	938	922
2012–2014		2590	2590	1286	1172	798	***663***	897	817	897	817	708	671

Note: All models include annual and a maximum of 598 census tract fixed effects. A maximum of 2000 tokens are used. Minimum values for each subperiod are indicated with bold italics.

DOI: 10.1002/jae

Table V. Repeat sales

		LS		FDR=0.1		CV		BC		Post		Het	
	(BASE)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
Panel A: In-sample RMSE													
2000–2014	0.433	***0.366***	0.377	0.368	0.382	0.367	0.378	0.382	0.392	0.375	0.385	0.387	0.395
2000–2007	0.224	***0.184***	0.192	0.193	0.200	0.188	0.195	0.199	0.205	0.195	0.200	0.204	0.210
2008–2014	0.474	***0.332***	0.363	0.368	0.387	0.345	0.366	0.379	0.405	0.362	0.386	0.387	0.413
2012–2014	0.450	***0.254***	0.282	0.404	0.433	0.313	0.333	0.362	0.404	0.339	0.376	0.367	0.403
Panel B: Out-of-sample RMSE													
2000–2014	0.431	0.374	0.385	0.376	0.390	***0.373***	0.384	0.384	0.393	0.379	0.389	0.388	0.396
2000–2007	0.223	0.196	0.233	0.197	0.227	***0.194***	0.200	0.200	0.206	***0.194***	0.200	0.205	0.210
2008–2014	0.472	0.388	0.516	0.387	0.495	***0.365***	0.387	0.385	0.412	0.371	0.395	0.389	0.417
2012–2014	0.449	0.485	0.630	0.426	0.489	***0.352***	0.378	0.369	0.414	0.356	0.390	0.376	0.409
Panel C: $\hat{\boldsymbol{Q}}$													
2000–2014		2021	2021	789	621	1536	1494	313	301	313	301	236	***231***
2000–2007		2013	1913	207	129	769	787	101	109	101	109	101	***98***
2008–2014		2013	1713	171	189	792	821	125	143	125	143	***98***	125
2012–2014		1999	1907	32	***25***	292	353	50	49	50	49	61	62

Note: All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values for each subperiod are indicated with bold italics.

DOI: 10.1002/jae

in Tables III and IV are summarized here. First, we find that in-sample and out-of-sample RMSEs for U2 and B2 are less than the BASE RMSE. Second, the out-of-sample RMSE for any model estimated using LS models is prone to spike up. Third, RMSEs from the BC and Post estimation procedures are less prone to spike up than other sparse estimation procedures. The online Appendix also displays robustness checks.

In order to determine the predictive power of the tokens across size segments, we stratify the data into quartiles based on each house's living area in square feet. In the online Appendix we present the results of a hedonic model for houses in the lower (Table A5) and upper (Table A6) quartiles of house size. Similar to the hedonic results in Table IV, we find that token augmented models clearly outperform the BASE model both in-sample and out-of-sample. We find that the tokens improve performance more for the smaller homes than the larger homes.

4.2. Estimation when $N < K$

The choice of 2000 tokens was made because it provided a rich set of tokens and permitted computation in a reasonable amount of time.[11] Results in the online Appendix (Table A7) compare performance when using 500, 1000, 2000 and 3000 possible tokens and indicate that as few as 100 tokens can improve in-sample and out-of-sample RMSE. For example, using 100 unigrams in U2 and estimating with the LS procedure reduces RMSE by a factor of 0.942.

It is of interest to compare the relative performance of the estimators when $N \ll K$ to the situation when $K \ll N$. Of course, least-squares procedures are not possible when $N \ll K$; however, the various LASSO procedures are possible in this setting. Ideally, it would be interesting to compare performance across multiple N and K where $N \ll K$ or $K \ll N$. One possible scheme is to completely saturate the model with all tokens observed in the data. That is, beginning with the entire set of N, we might increase K such that $N \ll K$. However, Table A1 indicates that the in-sample and out-of-sample performance of penalized procedures are comparable to the in-sample performance of least squares when $K = 3000$. Thus including the remaining tokens that are not included when $K = 3000$ will not significantly improve performance.

An alternative scheme is to begin with a small number of observations where $N \ll K$ and compare the performance of the estimators as N increases. This can be done by randomly selecting a small number of properties in the data and repeating. Instead, we compare the performance of the estimators within a given zip code as N increases. This scenario is perhaps more relevant for researchers, mortgage holders and practitioners interested in producing accurate valuation models for a relatively small geographic area. We carry this out using the following procedure.

1. For zip code z, randomly select 500 observations from the entire set of N_z observations in zip code z. Define this as the estimation set.
2. Collect the K_z tokens that appear in two or more of the estimation sets. Estimate the penalized procedures using the estimation set, the control variables and the K_z tokens, and calculate the fivefold out-of sample RMSE.
3. Randomly select 500 additional properties from zip code z, if possible, and add them to the estimation set.
4. Estimate the penalized procedures using the estimation set, the control variables and the K_z tokens, and calculate the fivefold out-of sample RMSE.
5. Repeat steps 3 and 4 until all N_z properties have been used.

[11] For example, when using a 2.5 GHz Intel Core i5 processor with 16 GB of memory and distributing the program over four cores, the results in Table IV required approximately 3 hours for 2000 tokens. When using 3000 tokens, the results required more than 18 hours for the results in each table. A significant portion of the run time was the LS procedure and the Post procedure. Details of the computing times are available from the authors upon request.

DOI: 10.1002/jae

The above procedure begins with 500 observations and increases the number of observations by 500 at each step. In order to compare the estimators to least squares, we calculate the fivefold out-of-sample RMSE for least squares using all N_z observations with the K_z tokens and the control variables. The out-of-sample RMSE for the penalized regressors are then scaled by the fivefold out-of-sample RMSE for least squares. When the ratio is less than 1, a LASSO variant has superior out-of-sample performance. By scaling in this way, the relative performance of the penalized procedures and least squares can be compared across multiple zip codes as N increases. For the results below, we refer to the N_z observations for zip code z as the full sample.

It is important to point out a practical problem inherent in working with binary regressors similar to those generated by the tokens. For fixed N, it is not always possible to increase K in a meaningful way. When N is small and there are a large number of tokens, the number of listings that contain a

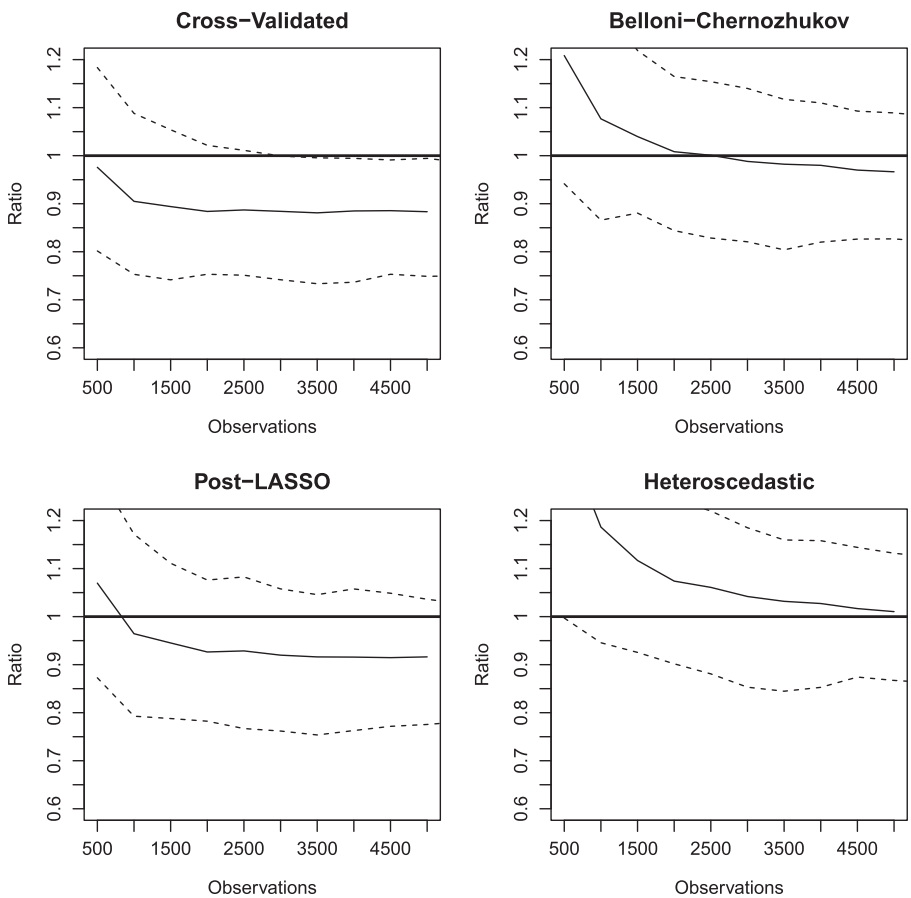

Figure 3. Unigram out-of-sample performance for $N < K$ and $K < N$. This figures displays the ratio of CV, BC, Post and Het LASSO out-of-sample RMSE to least-squares out-of-sample RMSE, within a zip code, for 89 zip codes. Least-squares out-of-sample RMSE is calculated using all observations in a given zip code. CV, BC, Post and Het LASSO out-of-sample RMSE are calculated using $500, \ldots, 5000$ observations where available. For each observation and zip code, the ratio of the CV, BC, Post and Het LASSO out-of-sample RMSE to the least-squares out-of-sample RMSE is calculated. A value of 1 indicates that CV, BC, Post and Het LASSO have out-of-sample RMSE equal to that of least squares when using all observations for a given zip code. The average out-of-sample RMSE across all zip codes within each observation size group is shown as a solid line. The 5% and 95% quantiles are shown with dashed lines

DOI: 10.1002/jae

given token can be quite small; alternatively, the number of non-zero entries in the associated regressor is quite small. In our data, we found that there were, on average, 796 (1176) unigrams (bigrams) appearing two or more times in a randomly selected set of 500 observations. Of course, it is possible to increase the number of tokens used, but this would necessarily create a situation where a token appears in only a single observation. Given the penalty function in the LASSO, it is not the case that this token can be used to perfectly predict the price of the associated observation as in least squares; however, given the infrequency of such a token within a small geographic area, estimating the value of the token does not appear to be of great importance.

Figure 3 summarizes the results of the above experiment using 89 zip codes in the data with unigram tokens. The cross-validated method outperforms least squares, on average, for all N. The Post-LASSO estimator begins to outperform least squares on average when the number of observations is 1000. As the average number of unigrams across the zip codes is 796, we interpret this result as the Post-LASSO unigram method having comparable performance to least squares on the full-sample when $K \sim N$. For 500 observations or $K \sim 2N$, the average cross-validated LASSO performance is better than full-sample least squares, and Post-LASSO performance is comparable to least-squares.

The online Appendix provides further results for $\hat{Q}$ (Figure A2). Similar to our previous results, cross-validation selects more variables than the other procedures. Interesting to note, although control variables are used, unigram tokens are selected when using 500 observations, providing evidence that text contains valuable information even when $N \ll K$. Results in the online Appendix confirm a similar performance when bigrams are used (Figures A3 and A4).

5. CONCLUSIONS

The linear hedonic model assumes that the price of a property is a linear combination of all of its attributes, both observed and unobserved. By including as many relevant variables as possible in the model, the researcher can minimize pricing error. Typically, the researcher only has a subset of variables available that they include in a hedonic model as either continuous or binary variables. As such, hedonic models are prone to an omitted variable bias if homebuyers value characteristics of a house that are not included in the subset of data employed in most real-estate studies. We show that the comments section of the MLS, which is populated by real-estate agents who arguably have the most local market knowledge and know what homebuyers value, provides information that improves the performance of both in-sample and out-of-sample pricing estimates.

We evaluate several penalized regression procedures that allow us to incorporate data in the MLS comments section in a hedonic model. Using data from the Atlanta MLS, we find that text data, in the form of unigram or bigram tokens, can be used alongside standard hedonic variables to improve both in-sample and out-of-sample RMSE when census tract fixed effects are utilized in both a standard OLS model and a penalized regression model. Our analysis evaluates the performance of penalized regression procedures across several time periods using two of the most common models in the real-estate literature: hedonic and repeat sales. We find that including textual information from the MLS comments section can decrease pricing errors by more than 25% relative to the models employed in most real estate studies.

Our results strongly suggest that future real-estate studies should incorporate the textual information available in the MLS comments section. The textual information is readily available for researchers using MLS data and we provide several procedures to harness the comments' predictive power. Additionally, studies with non-MLS data sources can also incorporate the textual information embedded in MLS comments as they are now readily available online on public websites such as Yahoo, Zillow and Trulia. Interestingly, Zillow states that they use MLS data feeds to validate the basic information they

DOI: 10.1002/jae

display for each property on their website.[12] Thus MLS data may result in a recalculation and change to Zillow's estimated value (Zestimate). However, to the best of our knowledge, although Zillow displays the MLS comment section on their website, they currently do not use the textual information in their value estimate.

ACKNOWLEDGEMENTS

We are grateful to Jarl Kallberg, Dongshin Kim, Crocker Liu, Feng Yao and three anonymous referees for their helpful comments and guidance. All errors are our own.

REFERENCES

Belloni A, Chernozhukov V. 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**(2): 521–547.

Belloni A, Chernozhukov V, Wang L. 2011. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98**(4): 791–806.

Belloni A, Chen D, Chernozhukov V, Hansen C. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**(6): 2369–2429.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**(1): 289–300.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**(4): 1165–1188.

Benson ED, Hansen JL, Schwartz AL Jr, Smersh GT. 1998. Pricing residential amenities: the value of a view. *Journal of Real Estate Finance and Economics* **16**(1): 55–73.

Bickel PJ, Ritov Y, Tsybakov AB. 2009. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**(4): 1705–1732.

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**: 993–1022.

Bollen J, Mao H, Zeng X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* **2**(1): 1–8.

Bond MT, Seiler VL, Seiler MJ. 2002. Residential real estate prices: a room with a view. *Journal of Real Estate Research* **23**(1–2): 129–138.

Brown GM, Pollakowski HO. 1977. Economic valuation of shoreline. *Review of Economics and Statistics* **59**(3): 272–278.

Bühlmann P, Van De Geer S. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer: Berlin.

Bunea F, Tsybakov A, Wegkamp M. 2007. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1**: 169–194.

Butler RV. 1980. Cross-sectional variation in the hedonic relationship for urban housing markets. *Journal of Regional Science* **20**(4): 439–453.

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. 2009. Reading tea leaves: how humans interpret topic models. In *Advances in Neural Information Processing Systems*. MIT Press: Cambridge, MA; 288–296.

Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456): 1348–1360.

Feng Y, Yu Y. 2013. Consistent cross-validation for tuning parameter selection in high- dimensional variable selection. arXiv preprint arXiv:1308.5390.

Fernandez C, Ley E, Steel MFJ. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**(5): 563–576.

Garcia D. 2013. Sentiment during recessions. *Journal of Finance* **68**(3): 1267–1300.

Gentzkow M, Shapiro J. 2010. What drives media slant? Evidence from US newspapers. *Econometrica* **78**(1): 35–71.

Goodwin K, Waller B, Weeks HS. 2014. The impact of broker vernacular in residential real estate. *Journal of Housing Research* **23**(2): 143–161.

[12] http://www.zillow.com/feeds/FeedsFAQ.htm#whatisfeed.

DOI: 10.1002/jae

Greenshtein E, Ritov Y. 2004. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**(6): 971–988.
Griffiths TL, Steyvers M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences USA* **101**(Suppl 1): 5228–5235.
Haag J, Rutherford R, Thomson T. 2000. Real estate agent remarks: help or hype? *Journal of Real Estate Research* **20**(1–2): 205–215.
Hill RC, Knight JR, Sirmans CF. 1997. Estimating capital asset price indexes. *Review of Economics and Statistics* **79**(2): 226–233.
Hofmann T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann: Burlington, MA; 289–296.
Kang H-B, Reichert AK. 1991. An empirical analysis of hedonic regression and grid- adjustment techniques in real estate appraisal. *Real Estate Economics* **19**(1): 70–91.
Knight K, Fu W. 2000. Asymptotics for lasso-type estimators. *Annals of Statistics* **28**(5): 1356–1378.
Leng C, Lin Y, Wahba G. 2006. A note on the Lasso and related procedures in model selection. *Statistica Sinica* **16**(4): 1273–1284.
Levitt SD, Syverson C. 2008. Market distortions when agents are better informed: the value of information in real estate transactions. *Review of Economics and Statistics* **90**(4): 599–611.
Loughran T, McDonald B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *Journal of Finance* **66**(1): 35–65.
Malpezzi S. 2003. Hedonic pricing models: a selective and applied review. In *Housing Economics and Public Policy*, O'Sullivan A, Gibb K (eds). Wiley: Chichester; 67–89.
Mayer CJ. 1998. Assessing the performance of real estate auctions. *Real Estate Economics* **26**(1): 41–66.
Meinshausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the LASSO. *Annals of Statistics* **34**(3): 1436–1462.
Mitra T, Gilbert E. 2014. The language that gets people to give: phrases that predict success on Kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM: New York; 49–61.
Paterson RW, Boyle KJ. 2002. Out of sight, out of mind? using GIS to incorporate visibility in hedonic property value models. *Land Economics* **78**(3): 417–425.
Pryce G, Oates S. 2008. Rhetoric in the language of real estate marketing. *Housing Studies* **23**(2): 319–348.
Rosen S. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* **82**(1): 34–55.
Rouwendal J, Van Marwijk R, Levkovich O. 2016. The value of proximity to water in residential areas. *Real Estate Economics*. (forthcoming) DOI: 10.1111/1540-6229.12143.
Sala-i Martin XX. 1997. I just ran two million regressions. *American Economic Review* **87**(2): 178–183.
Song Y, Knaap G-J. 2003. New urbanism and housing values: a disaggregate assessment. *Journal of Urban Economics* **54**(2): 218–238.
Soyeh KW, Wiley JA, Johnson KH. 2014. Do buyer incentives work for houses during a real estate downturn? *Journal of Real Estate Finance and Economics* **48**(2): 380–396.
Tetlock PC. 2007. Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance* **62**(3): 1139–1168.
Taddy M. 2013a. Measuring political sentiment on twitter: factor optimal design for multinomial inverse regression. *Technometrics* **55**(4): 415–425.
Taddy M. 2013b. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association* **108**(503): 755–770.
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**(1): 267–288.
Tu CC, Eppli MJ. 1999. Valuing new urbanism: the case of Kentlands. *Real Estate Economics* **27**(3): 425–451.
Varian HR. 2014. Big data: new tricks for econometrics. *Journal of Economic Perspectives* **28**(2): 3–27.
Wainwright MJ. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using- constrained quadratic programming (LASSO). *IEEE Transactions on Information Theory* **55**(5): 2183–2202.
Wallach HM, Murray I, Salakhutdinov R, Mimno D. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM: New York; 1105–1112.
Zhao P, Yu B. 2006. On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**: 2541–2563.
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.
Zou H, Hastie T, Tibshirani R. 2007. On the 'degrees of freedom' of the lasso. *Annals of Statistics* **35**(5): 2173–2192.

DOI: 10.1002/jae