

# Are Solar Panels Commodities? A Bayesian Hierarchical Approach to Detecting Quality Differences and Asymmetric Information.

Johannes Mauritzen\*

April 10, 2019

## Abstract

Solar panels should not be considered commodities. Considerable quality differences, as measured directly by degradation of production over time, are found between manufacturers. I test two implications from the theory of asymmetric information of quality and find: 1.) Solar power systems with high-information third-party owners display higher quality than host-owned systems. 2.) Furthermore, with a 85% probability, the price of solar panels that are owned by high-information owners are more highly correlated to quality. Methodologically, the article demonstrates a novel application of Bayesian hierarchical regression models that are increasingly popular in operations research and the decision sciences.

**Keywords**— Solar power, commodities, asymmetric information, quality, Bayesian hierarchical model

---

\*BI Norwegian Business School, johannes.mauritzen@bi.no, <http://jmaurit.github.io>

# 1 Introduction

Solar panels have been referred to colloquially as being commodities.<sup>1</sup> This is often in reference to the high degree of competition, especially from Chinese manufacturers, that has dramatically driven down prices of solar panels in the last decade. This has made solar power price competitive without subsidies in many sunny locations. But it has also meant upheaval in the panel manufacturing industry with cheaper Chinese produced panels replacing panels produced in Europe, North America and Japan. Most recently, this led to the imposition of trade tariffs on Chinese produced solar panels by the US government.

More formally, a commodity is generally defined by a certain degree of fungibility. That is, the good from a certain producer will be largely interchangeable with the good of another producer. Already in 1873 the German economist A.E.F. Schäffle distinguished between a popular understanding of a commodity as a material input in manufacturing, to the scientific understanding as "exchangeable material goods" (Menger et al., 2007). Despite the simplicity of this definition, Gordon et al. (1999) show the difficulty in practice of determining whether a product is in fact a commodity. The root problem is that a good can be differentiated in a host of ways, some of which are difficult to observe at the time of transaction.

Determining whether a product is a commodity is of more than just theoretical importance. Historically, the commodification of goods was a necessary condition for the establishment of sophisticated financial markets for those goods. A well-documented example: The invention of grain elevators for mixing and standardising grain was a necessary precursor to the establishment of a futures markets with receipts of delivery traded on the Chicago Board of Trade (Cronon, 1992). The abstraction between producer and end purchaser of goods was necessary for the functioning of these markets. A trader could fulfil a promise to deliver grain of a certain grade by simply buying it on the market or with cash settlement, rather than having produced it oneself.

An important differentiating characteristic between solar panels is the quality of these solar panels. This can be measured as both the failure rate of these panels as well as the degree of degradation of the panels over time. If there were significant differences in quality between manufacturers, then this has implications for the structure and development of the solar panel market. Manufacturers with superior quality could differentiate themselves and extract a premium in the market. A reputation for quality among established manufacturers could also act as a barrier to entry, thereby reducing long-term pricing pressure.

From the perspective of a purchasing manager responsible for procuring solar panels, whether panels can be considered a commodity is of great importance. If panels are known to be commodities with standardised quality and similar pricing, then the decision for purchasing manager is substantially simplified. They need only to decide on the observable form factor, for example dimension and capacity. In addition, purchasing managers responsible for large projects with long lead times could make use of forward markets to hedge price movements. On the other hand, if solar panels are known to vary significantly by quality, then the decision becomes substantially more complex and costly. Purchasing managers would need to weigh the price of panels against expected quality. They may also need to make use of specialised engineering

---

<sup>1</sup>See for example <https://www.greentechmedia.com/articles/read/are-pv-modules-a-commodity#gs.lx3tMbY>

firms that test and certify panel quality. Since the purchasing decision is likely to be linked to a specific manufacturer, the purchasing manager is subject to risks such as bankruptcy or other supply disruption associated with that firm.

Furthermore, if significant quality differences between manufacturers exists, then the characteristics of solar panel investments suggest the possibility of asymmetric information of quality in the market. This is important in practice, since the presence of information asymmetry in a market tends to lead to the persistence of low quality suppliers.

The market for rooftop solar panels can be expected to be particularly vulnerable to issues of asymmetric information on quality. Solar panels can be characterised as an “experience” good, where an investor/consumer needs to learn about the quality through use (Caulkins et al., 2017). In particular, poor quality panels will tend to show a higher degradation of output over time than high quality panels. Even then, solar panel owners may find it difficult to measure the degradation as it can happen gradually, over many years.

More so, solar panel systems are expected to last at least 20 years, thus for all practical purposes, their purchase can be considered a one-shot investment. This eliminates repeat buying as a mechanism for ensuring quality. In the literature, warranties are often suggested as a strong signal of quality. However, warranties may be a relatively weak assurance of quality in the market for solar panels as both contractors and manufacturers are relatively new, tend to be heavily indebted and some have recently shown a tendency to go bankrupt.

Given the difficulty of judging quality and lack of market mechanisms to signal quality, the established economic theory on the subject would suggest that the market may tend to provide low-quality panels at relatively undifferentiated prices (Tirole, 1988).

Over the period from 2010 to 2016, many of the larger solar contractors moved to leasing solar power systems to home- and business-owners (Mauritzen, 2017). These contractors, who buy in scale, have the resources and expertise to judge quality in solar panels. They can take steps such as getting panels tested at specialized engineering firms and visiting manufacturing sites to ensure quality of suppliers. This descriptive understanding of the market suggests that it is possible to divide the market into two types of solar panel owners: high-information owners who lease the solar panel systems to individual home owners, businesses and other organizations, and low-information owners who buy and own the solar panel systems directly.

From this informal model of information asymmetry in the market for solar panels, a testable implication arises. The quality of panels, as measured by degradation of output over time, would be predicted to be higher in the systems owned by the high-information owners (third-party-owner) who can discern between low and high quality solar panels. Conversely, the basic theory of asymmetric information suggests that lower quality panels will tend to push out higher quality panels in the presence of low-information buyers who cannot discern quality, and therefor make purchasing decision solely on price.

A related prediction from the theory of asymmetric information of quality is that the cost of solar panel systems should have a higher degree of correlation with quality when the owner is a high-information type.

Because a high-information owner can, by definition, differentiate between high- and low-quality panels.

In this paper, I use a data set of California solar power systems that includes monthly production data from over 3000 solar panel systems. Where many studies of asymmetric information of quality use proxies for quality, I measure it directly as the average degradation of production over time.

In summary, I have three nested research hypothesis that I wish to test empirically.

1. Solar panels are not pure commodities. Significant quality differences can be detected between manufacturers of panels.
2. Given that quality differences exist between manufacturers, high information owners (third-party owners) will purchase higher quality panels than low-information owners (host owners).
3. Price and quality will be more highly correlated in solar panel systems owned by high information owners.

The data has a natural hierarchy of between 36 and 60 monthly observations within the 3000-plus solar panel systems, which in turn can be grouped according to relevant characteristics such as whether they are host-owned or third-party owned. Because these groupings display considerable heterogeneity in sample size, mean values and variance, it is natural to model the relevant structure of the data directly. For relatively parsimonious models with one or two hierarchies and few covariates, I can use standard linear mixed effects models (LMEM) estimated by maximum likelihood. However, for richer models with all relevant hierarchies, maximum likelihood becomes impractical.

As a solution, I use a Bayesian hierarchical<sup>2</sup> model estimated using Markov Chain Monte Carlo (MCMC) simulation techniques. Bayesian hierarchical models have long been one of the primary tools in statistical decision theory (Berger, 1985) and this form of modelling is increasingly common in operations research and management science. One of the main advantages is superior out-of-sample predictive properties with data with distinct groupings. Recent advances in computation have also allowed for the construction and simulation of increasingly complex models. In particular, I use the Stan probabilistic programming language (Stan Development Team, 2014) that makes use of Hamiltonian Markov Chain Monte Carlo (MCMC) (MacKay, 2003) and a No-U-Turn sampler (Homan and Gelman, 2014). Unlike earlier MCMC simulation routines, notably Gibbs sampler, Hamiltonian MCMC is able to reach convergence even with models, like the one I present in this article, with multiple hierarchies. Importantly, with the Hamiltonian MCMC routine, it is not necessary to use conjugate distributions, opening up to a wider variety of choice of prior distributions (Kruschke, 2014).

In part due to increasing computational tractability, Bayesian hierarchical models have recently become a common and useful tool in the operations research field. A common application is to incorporate expert opinion into quantitative decision models (Rufo et al., 2010; Szwed et al., 2006; Utkin, 2006). In this application, the Bayesian hierarchical model is flexible enough to aggregate the information of the expert opinions while also being able to explicitly handle and propagate the uncertainty in these opinions. Hierarchical models have also been used to identify default risk among lenders and improve credit scoring models (Liu et al., 2015; Bhattacharya et al., 2019). In this context, the hierarchical Bayesian models are

---

<sup>2</sup>Sometimes also called multilevel model

found to provide superior out-of-sample prediction for the censored sample of the data. Other operations research related applications include a fund manager’s portfolio selection problem (Soyer and Tanyeri, 2006; Bodnar et al., 2017) and conjoint analysis in operations management (Karniouchina et al., 2009; Gensler et al., 2012). A commonality among all these applications is the need to aggregate information from heterogenous groups without over-fitting the data with many fixed effects.

Despite the usefulness of hierarchical Bayesian models, examples of hierarchical models in the economics and finance literature are sparse. Dehejia (2003) uses a Bayesian hierarchical model to evaluate the effectiveness at multiple programs at different sites. By incorporating “predictive uncertainty” by way of a hierarchical model, the impact of the sites, which otherwise is estimated to be significant with a fixed effects model, disappears. Meager (2018) uses a hierarchical model to aggregate studies of micro-credit expansion in developing countries in order to estimate an overall treatment effect with “external validity,” that is the degree to which an estimated treatment effect generalizes to other populations. Jones and Shanken (2005) uses a hierarchical model to study returns of mutual funds, arguing that estimates of long-run returns of any single fund should be informed, by way of a higher-level distribution, of the returns of the population of other mutual funds. In a study with parallels to my own, Korteweg and Sorensen (2017) uses a Bayesian hierarchical model to study persistence of returns in the private equity sector both between and within firms, where persistence is measured by excess variation between firms as captured by firm-level parameters (or random effects) that are drawn from a higher-level distribution. If returns are thought of as the converse of production degradation, then the existence of persistence of returns can be thought of as the converse of quality differences.

The property of hierarchical models that allows for improved predictive performance with the presence of groupings in the data is regularization, or alternatively called partial pooling (Gelman et al., 2013). As opposed to fixed effects estimation where a fixed effects coefficient is determined exclusively with data from within a particular group, in a hierarchical model each of the group parameters are regularized towards a higher-level parameter estimated with the full data in proportion to the relative number of data points in a particular group.

The linear mixed effects models estimated by maximum likelihood that I use in the first section of this article also have the regularization property. Bayesian estimation provides several theoretic advantages, such as correct inference in small samples and generalization to distributions other than the normal. However, for my purposes, the main benefit of using a Bayesian methodology is practical. I can fit a rich model with several hierarchies and many thousands of parameters, that simply cannot be handled by maximum likelihood procedures.

The issue of how information asymmetries of quality affects market outcomes was first introduced by Akerlof (1970). Ch. 2 of Tirole (1988) and accompanying citation list provides a good overview of the theoretical foundations of this topic.

Early theoretical work of particular relevance to this paper includes Chan and Leland (1982), whom model the situation where information on quality is costly rather than directly unavailable and where quality is endogenous to the seller—in other words the seller can choose the level of quality of its product.

Under these assumptions, the authors predict several equilibrium—essentially markets for both high- and low-quality products depending on the extent to which the buyer is informed.

Reputation, or lack thereof, plays a particularly important role in this market. Most solar installers and panel manufacturers are relatively new firms, especially the increasingly dominant Chinese solar panel producers. Shapiro (1982) explores the role of reputation and finds that because reputation is effective only with a lag, the "steady-state" quality level chosen by a seller should be lower than in the scenario with perfect information.

An expanding empirical literature exists on asymmetric information of quality, especially in the marketing, finance and accounting fields. In accounting, the issue of firm quality and voluntary disclosure in regulatory filings as a signal has received broad attention by empirical researchers. Healy and Palepu (2001) review the early literature. Empirical studies in finance have, for example, looked at issues of adverse-selection and quality in Initial Public Offerings (Michaely and Shaw, 1994), small business lending (Petersen and Rajan, 1994), and in a recent application for consumer credit markets (Dobbie and Skiba, 2013).

The literature in the marketing field is also relevant to this paper, as it is particularly concerned with the effects of information asymmetry on purchases of consumer durables. The effects of signalling on quality, usually in the form of advertising, is particularly important in the field. Kirmani and Rao (2000) provide an overview of this literature.

The use of leasing, discussed in this article, can be seen as a form of costly signalling about quality. However leasing plays more than just an informational role since it also transfers much of the risk of quality to the contractor. Traditionally, energy generation has been an activity undertaken by large, specialized firms with specialized knowledge and resources. The role of information asymmetry in investment and purchase decisions has likely been seen as a secondary issue, and to my knowledge, no literature exists on the role of such informational issues in the industry.

However, the growth of distributed energy technologies, like solar panels, has added new focus on informational and behavioral issues related to energy investments by homeowners and small businesses. In turn, a growing literature is forming around solar power investment behavior of homeowners and small businesses.

For example Dastrup et al. (2012) argue that solar panels cannot be considered a pure investment good, but are also bundled as a type of green conspicuous consumption, showing evidence for a "solar price premium" in homes with solar panels. Bollinger and Gillingham (2012) study the role of peer effects in solar power adoption. They find evidence that the adoption of solar panels by homeowners in a certain zip-code will increase the probability that other households in that zip-code will install solar panels. A focus in the operations research field has been on designing systems to manage and aggregate such distributed energy resources on behalf of un-sophisticated owners (Velik and Nicolay, 2016; Le Cadre et al., 2019).

Despite the growth of the literature on investments in distributed energy, to my knowledge this is the first paper to identify and directly test for quality differences and related role of information asymmetry

in the expanding market. The article has implications for an understanding of the market for distributed energy technologies, subsidy policies and regulation of the market.

## 2 California Solar Initiative, Data

The data used in this article is an intersection of two data-sets from the California Solar Initiative (CSI). The California Solar Initiative is a state-wide program that gives incentives for installing grid-connected solar panel systems. The incentives are based on performance. For smaller systems, typical of most residential and small business installations, this incentive could be given in the form of an upfront pay-out based on the capacity and the expected performance of the system.

Larger systems were required to accept a performance-based incentive based on actual production over 5 years (60 months). From the beginning of the program in 2007 this was defined as those over 100 kW, but was lowered to 50 kW from January 2008 and 30 kW from January 2010. Solar panel installations of all sizes have the option of getting an incentive based on actual performance.

CSI provides data on all grid-connected solar panel systems installed in California since January 2007. In addition, CSI provides another data set of monthly production data from the solar panel systems that received production incentives. The data is openly available on the website of CSI.<sup>3</sup>

The installation data can be matched with the production data for those installations receiving a production subsidy. I include only those installations that have been producing for at least 3 years (36 monthly observations), where the maximum number of observations in this data is 5 years (60 months) of production data.

I removed systems from the data that may have had reporting errors. For example, I removed systems where production was reported to be higher than what would be theoretically possible from a solar power system with a given nameplate capacity. I also removed systems that reported 0 production in a period. This could also be an indication of quality if zero production indicates a malfunction in the system. However, it is not possible to identify which zero observations reflect malfunctions versus reporting issues or other non-quality issues.

Table 1 gives an overview of the variables in the data that are used in the analysis. Notice that variables can vary by monthly observation or by solar power system. I will take account of this natural data hierarchy in my modelling. A total of 171,027 monthly observations are used in the analysis. These observations come from 3,145 unique solar power systems. From these solar panel systems, panels from a total of 107 unique manufacturers are used.

Production is measured in kilowatt hours (kWh). In the analysis, production data is de-scaled, as shown in equation 1. For production data for each system,  $s$ , the production data observations for each month,  $i$ , is demeaned and then divided by the standard deviation of the production data for that solar power system. Figure 1 gives a visualization of the transformed production data. The red line shows the average degradation over time of all the systems. The strong seasonality of production is also clearly visible.

---

<sup>3</sup>[http://www.californiasolarstatistics.ca.gov/current\\_data\\_files/](http://www.californiasolarstatistics.ca.gov/current_data_files/)

Table 1: Summary of data. Information is provided on whether the variables are at the observation (system-month) level or at the system level, Whether the variable values are real, integer or categorical in nature and the mean value of observations is also provided if relevant. The parenthesis after the sector categories in the footnote indicate the number of systems for each sector.

Variable	Label	Level	Values	Mean Value
Production (kWh)	prod_kwh	Observation	Real	35,908
Months of operation	months_operation	Observation	Integer	–
Month of year	month	Observation	Categ. <sup>1</sup>	–
Cost per kw	cost_per_kw	System	Real	7,026
Capacity rating (kW)	csi_rating	System	Real	239
First year of production	first_prod_year	System	Integer	2011
Third party owned	lease	System	Categ. <sup>2</sup>	35% leased
Owner sector	own_sector_id	System	Categ. <sup>3</sup>	–
Host county	county_id	System	Categ.	–
# Observations		171,027		
# Solar panel systems		3,145		
# Module manufacturers		107		
Min. months of data		36		
Max. months of data		60		

<sup>1</sup>January - December. <sup>2</sup>Leased 1, Owned 0. <sup>3</sup> Residential (497), Commercial (1759), Governmental (777), Non-profit (112).

Other continuous variables are transformed in a similar manner, with observation-level variables de-scaled at the system level, and system-level variables de-scaled across systems. These transformations avoid confusion that may come when interpreting coefficients on variables with different units and scales. There are also technical reasons for the transformations relating to the MCMC routine. Convergence in the MCMC routine is aided when all variables are on the same scale. Setting appropriate weakly informative prior distributions on the parameters is also simplified when continuous variables are free from units.

$$prod_{s[i]} = \frac{prod\_kwh_{s[i]} - mean(prod\_kwh_{s[i]})}{std(prod\_kwh_{s[i]})} \quad (1)$$

Month-of-year dummy variables are used to control for the seasonality in the data. The capacity of the solar panel system is the actual maximum capacity of a solar power system as reported to and audited by the CSI. This can be somewhat lower than the total nameplate capacity of the solar panels, and can depend on not only the technical capacity of the panels, but also on supporting equipment, such as the inverter, and factors such as the angle of the roof and orientation of the solar panel system.

The cost per kilowatt (\$/kW) variable is the reported total cost of the solar panel system divided by the CSI capacity rating. For solar panel systems that are sold outright, the total cost is simply the transaction price of the entire system. For systems with third party owners, the total cost is an estimate reported by the system owner based on the sum of component costs, installation costs and a mark-up. I will discuss how this may affect the results.

In the analysis I distinguish between solar panels that are host-owned and those that are third-party owned, which I refer to as being leased. What I refer to as leased can however take two main forms. The most common is a Power Purchase Agreement (PPA) where the host agrees to buy the electricity from the solar panel system for an agreed upon price, (which may increase at an agreed upon rate over time).



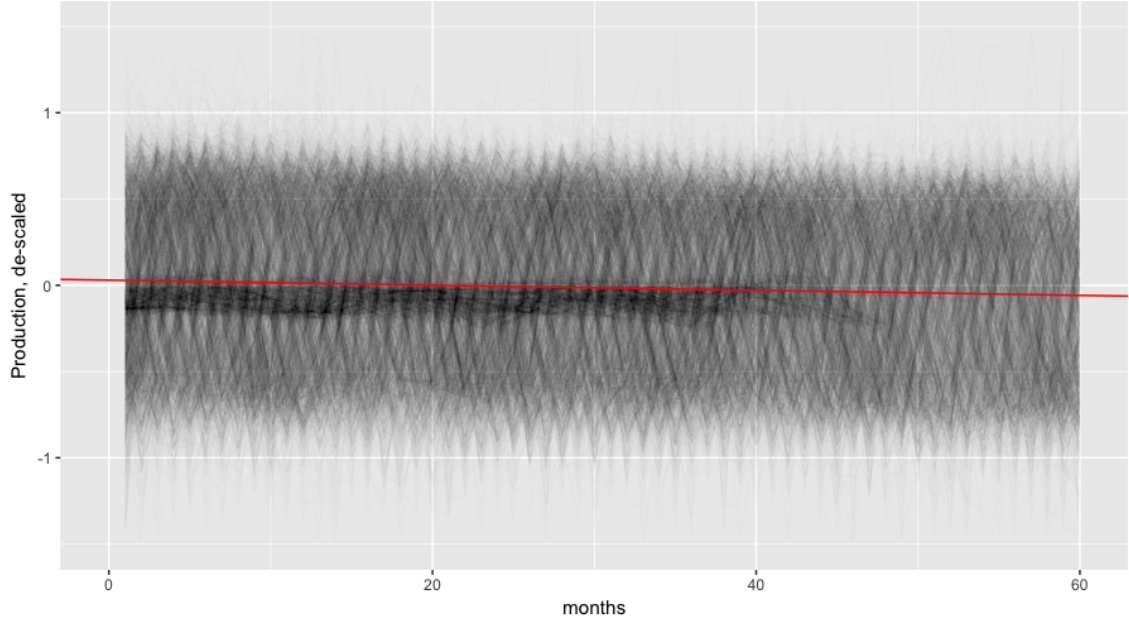


Figure 1: The figure shows the de-scaled production over time from the sample of solar panel systems. The red line represents the average degradation over time for all panels.

Alternatively, a host may pay a flat monthly fee for the electricity produced by the solar panel.

Finally, among the 3145 plus solar power systems, 497 had residential owners, 1759 had commercial owners, 777 had governmental owners and 112 had non-profit owners. A portion of the commercial owners operate as third-party-owners, with hosts in the other sectors. Alternatively, I could include host sector in my analysis: 1241 solar panel systems had a commercial host, 501 had a residential host, 1199 had a governmental host and 204 had a non-profit host. However, in a study of asymmetric information, it makes most sense to focus on the sector of the owner.

### 3 Linear Mixed Effects Models Estimated by Maximum Likelihood

#### 3.1 Testing for differences between module manufacturer

Following the notation of Bates et al. (2015), the conditional distribution of the response variables,  $y$  in a linear mixed effects model can be written compactly as in equation 3. Here  $\beta$  is a  $p$ -dimensional coefficient vector (or “fixed effects”) of the  $n \times p$  matrix  $\mathbf{X}$  of  $(p-1)$  explanatory variables plus intercept term.  $\mathbf{Z}$  is an  $n \times q$  model matrix of the variables whose groups are to be modelled directly with corresponding vector of coefficients, or “random effects”  $\mathbf{B}$  that are fixed at values  $\mathbf{b}$ .<sup>4</sup> Here  $q$  is dependent on the number of terms,  $k$ , that are allowed to vary by group and the number of groups,  $l$ , for each term:  $q = \sum_i^k q_i = \sum_i^k l_i p_i$ .  $\mathbf{W}$  is a diagonal matrix of known prior weights.  $B$  is distributed as multinomial normal with zero mean and

<sup>4</sup>The use of “random” refers to the fact that these groups are assumed to be specific to the sample. Many applications of mixed effects model thus treat the varying group coefficients as group-level error-terms. Interpreting coefficients as group varying coefficients or as group error terms is mathematically equivalent.

Symbol	Definition
$prod_i$	Production from a given system in a given month (observation i), descaled
<b>month</b>	Vector of month dummy variables
$months\_operation_i$	Number of months of operation of a system at observation i, descaled
$\mathbf{a}_s$	Intercept parameter for a system, s
$\mathbf{b}_s$	Slope parameter for a system, s
$\sigma^2$	Variance of the production data
$\mu^a$	Mean of the system intercept terms, $a_s$
$\mu^b$	Mean of the system slope terms, $b_s$
$\sigma^{2,a}$	Variance of the system intercept terms, $a_s$
$\sigma^{2,b}$	Variance of the system slope terms, $b_s$
$\rho$	Correlation coefficients between intercept and slope terms

Table 2: Definitions from the restricted model presented in equation 4.

where  $\Sigma$  is variance-covariance matrix.

$$(y|\mathbf{B} = \mathbf{b}) \sim \mathcal{N}(\mathbf{X}\beta + \mathbf{Zb}, \sigma^2 \mathbf{W}^{-1}) \quad (2)$$

$$\mathbf{B} \sim \mathcal{N}(\mathbf{0}, \Sigma) \quad (3)$$

Models of this form can be solved efficiently using a penalized maximum likelihood routine. For further details of both the model and routine, I refer to Bates et al. (2015).

To test whether I can detect quality differences between module manufacturers, I need to first estimate a restricted comparison model. I write the restricted comparison model explicitly in equation 4. Symbols are defined in table 2.

In this model, the transformed production variable,  $prod\_scaled_i$  for observation  $i$  is distributed normally with the mean modelled as a linear function of an intercept term  $a_s$ , a slope term,  $b_s$  on the transformed months of operation variable  $months\_operation\_scaled$ , and a set of dummy variables representing the month of year, **month**. Notice that the intercept and slope are indexed by  $s$ , indicating that they are allowed to vary by system. Thus, in this model formulation, I take into account the groupings by system, explicitly modelling varying relative initial production values (intercepts) and degradation over time (slope). Moving up a level, the estimated  $a$  and  $b$  parameters are in turn modelled as multivariate normal distributions.

$$prod_i \sim N(\mathbf{month} + a_s + b_s months\_operation_i, \sigma^2) \quad (4)$$

$$\begin{pmatrix} a_s \\ b_s \end{pmatrix} \sim N \left( \begin{pmatrix} \mu^a \\ \mu^b \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho \sigma^a \sigma^b \\ \rho \sigma^a \sigma^b & \sigma^{2,b} \end{pmatrix} \right)$$

The inclusion of the matrix of month dummies, **month**, controls for seasonal variation across all systems. This is a simplified way to model seasonality, but because the data is itself at a monthly frequency, and a more direct modelling of seasonality is not of direct importance, this should suffice. As long as unmodelled variation in seasonality is not correlated with the long-run slope of production, the point estimates for the slope should be unbiased.

Symbol	Definition
$\mu_m^a$	Mean of the system intercept terms, $a_s$ by manufacturer, $m$
$\mu_m^b$	Mean of the system slope terms, $b_s$ by manufacturer, $m$
$\mu_l^a$	Mean of the system intercept terms, $a_s$ by ownership type, $l$ (leased/host-owned)
$\mu_l^b$	Mean of the system slope terms, $b_s$ by ownership type, $l$ (leased/host-owned)
$\theta^a$	Higher level (meta) mean of the distribution of grouped intercept means, $\mu_m^a$
$\theta^b$	Higher level (meta) mean of the distribution of grouped slope means, $\mu_m^b$
$\sigma^{2,\mu_m^a}$	Higher level (meta) variance of the distribution of grouped intercept means, $\mu_m^a$
$\sigma^{2,\mu_m^b}$	Higher level (meta) variance of the distribution of grouped slope means, $\mu_m^b$

Table 3: Definitions from the unrestricted models presented in equation 5, with means that vary by manufacturer, and in equation 6, with means that vary by owner type (leased or host-owned). See table 2 for symbols that are identical in both the restricted and unrestricted models.

To test whether quality varies among module manufacturers, I simply add panel manufacturer as an extra hierarchy to the model. In this model, the observation-level regression is identical to the restricted comparison model as in equations 4. New symbols in this model are defined in table 3. The system-level intercepts,  $a$ , and slopes,  $b$ , are also modelled as multivariate normal distributions. But the mean values of the multivariate normal distribution,  $\mu^a$  and  $\mu^b$ , are now indexed by  $m$ , as shown in equation 5, representing that they are allowed to vary by module manufacturer.

$$\begin{aligned}
\begin{pmatrix} a \\ b \end{pmatrix} &\sim N \left( \begin{pmatrix} \mu_m^a \\ \mu_m^b \end{pmatrix}, \begin{pmatrix} \sigma^{2,a} & \rho\sigma^a\sigma^b \\ \rho\sigma^a\sigma^b & \sigma^{2,b} \end{pmatrix} \right) \\
\mu_m^a &\sim N(\theta^a, \sigma^{2,\mu_m^a}) \\
\mu_m^b &\sim N(\theta^b, \sigma^{2,\mu_m^b})
\end{aligned} \tag{5}$$

If there is no significant variation between the module manufacturer groupings, this model collapses into the restricted comparison model, which is the null hypothesis. Figure 2 gives a visual summary of the system-level slope terms ordered from lowest to highest. The lines represent approximate 95% confidence intervals around the point estimates. One anomaly is that a handful of the systems have slopes that are estimated to be positive. This is problematic if we wish to interpret the slope coefficient as degradation over time. I address this in the full Bayesian model without resorting to excluding data.

Figure 3 gives a summary of the estimated manufacturer-level mean slopes. Again, the lines represent 95% confidence intervals around the point estimates. Visually, we can see substantial variation in quality between the manufacturer groups. We can formally test the hypothesis that manufacturer groupings improve the explanatory power of the model with a comparison to the restricted model with a likelihood ratio test. The results are shown in table 4. The information criteria measures (AIC, BIC) are both lower, indicating better predictive power of the model with manufacturer groupings. The Chi-square statistic indicates that the model with manufacturer groupings substantially improves model performance.

The first result, of which the evidence is strong, is that solar panel systems vary substantially in quality between module manufacturers. Given that module quality is a salient differentiating characteristic between panels, it is then difficult to view solar panels as commodities.

Figure 2: Approximate 95% confidence intervals around maximum likelihood point estimates of system-level slope terms,  $b_i$ , ordered from lowest to highest. Visually, we can see significant differences in the average degradation over time across solar power systems.

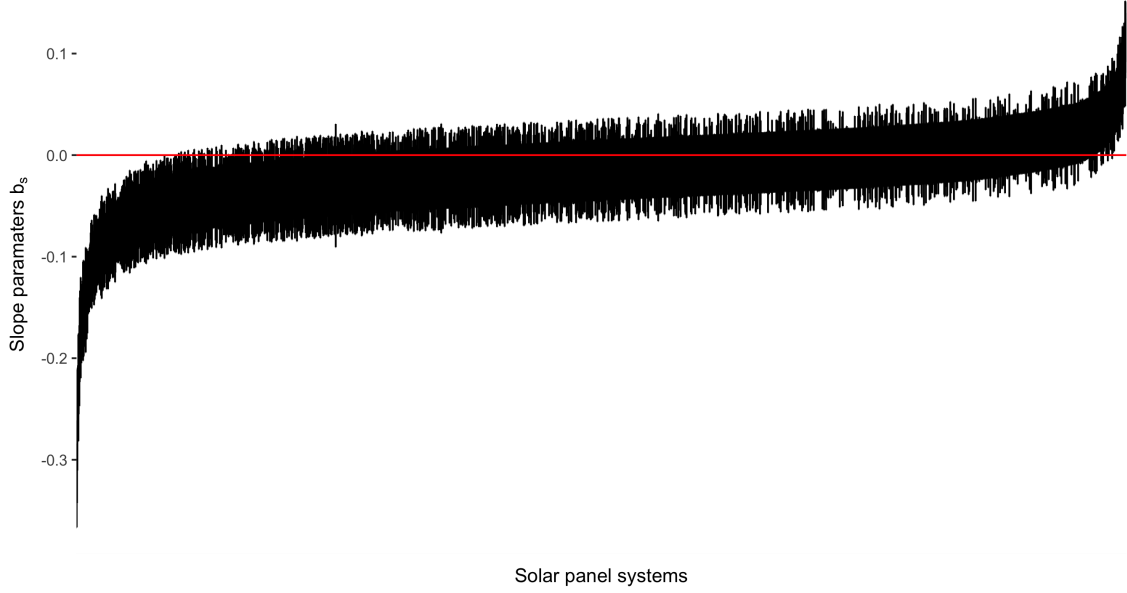
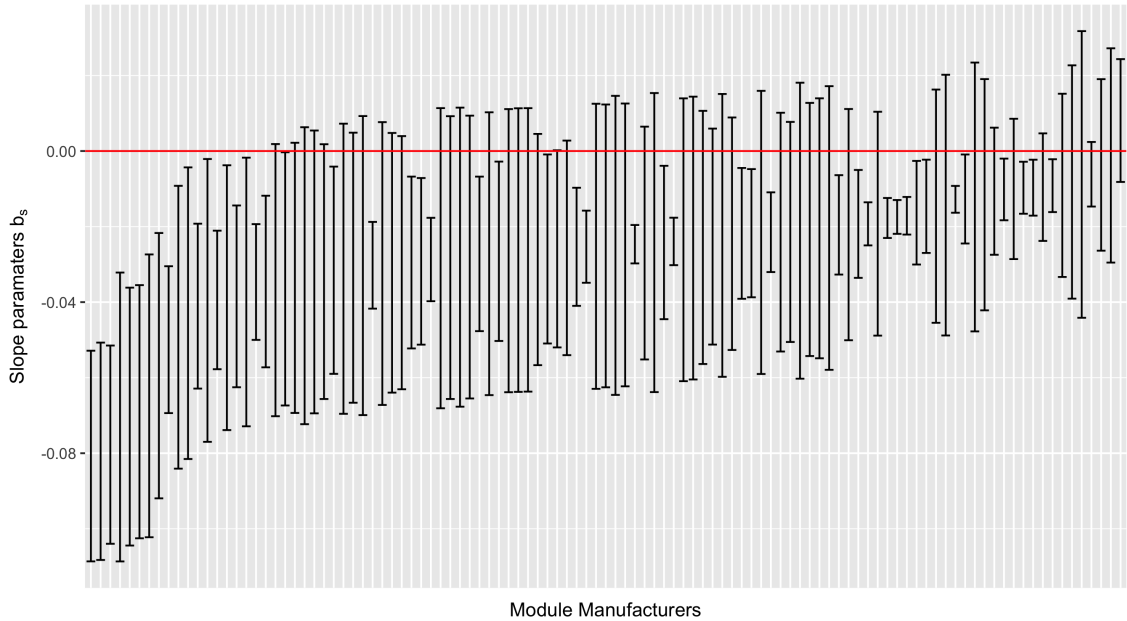


Figure 3: Approximate 95% confidence intervals around maximum likelihood estimators of manufacturer-level mean slope terms  $\mu_b$ , ordered from lowest to highest. These are the means of the higher-level normal distributions that vary by manufacturer that the system-level parameters are assumed to be drawn from. The visual impression is of substantial variation between manufacturer groupings.



Model	DF	AIC	BIC	logLik	deviance	Chisq	Chi df	$Pr(> Chisq)$
Restricted	17	8805	8976	-4385	8771	—	—	—
Grouped Manuf.	20	8430	8631	-4195	8390	381	3	2.6e-82

Table 4: A likelihood ratio test indicates that the addition of panel manufacturer as a hierarchy significantly improves the predictive power of the model. Akaike Information Criterion (AIC) and Bayesian Information Criterion are lower in the model with manufacturer groupings, indicating better out-of-sample predictive fit.

### 3.2 Testing for asymmetric information: differences in quality between leased and host-owned solar power systems

Provided the evidence that solar panel systems are not commodities and that substantial differences in quality do exist between manufacturers, we can now move on to the testable implication that comes from a combination of standard economic theory of asymmetric information and a descriptive understanding of the market that suggests solar panels could be subject to issues of asymmetric information of quality.

From our informal model of asymmetric information where we see host-owners of solar systems as being low-information types, while owners of leased panels as being high-information types, we would predict that on average leased panels will display higher quality over time. To test this directly with the data, I only need to do a slight modification of the model in the previous subsection. Instead of grouping the system-level coefficients by manufacturer, I now group into whether the solar power system was leased or not, as indexed by  $l$ . We can then compare the average slopes between these groups. For convenience the model is presented in full in equations 6.

$$\begin{aligned}
prod_i &\sim N(\mathbf{month} + a_s + b_s months\_operation_i, \sigma^2) \\
\begin{pmatrix} a \\ b \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_l^a \\ \mu_l^b \end{pmatrix}, \begin{pmatrix} \sigma^{2,a} & \rho\sigma^a\sigma^b \\ \rho\sigma^a\sigma^b & \sigma^{2,b} \end{pmatrix}\right) \\
\mu^a &\sim N(\theta^a, \sigma^{2,\mu^a}) \\
\mu^b &\sim N(\theta^b, \sigma^{2,\mu^b})
\end{aligned} \tag{6}$$

A visual summary of the results is shown in figure 4. Again, the vertical lines represent approximate 95% confidence intervals over the point estimates for the system-level slope terms,  $b_s$ . But here the slope estimates are grouped into host-owned (left) and leased (right). The estimated group means are shown as horizontal lines. At the scale of the figure, the effect appears small, but the average slopes are estimated to be significantly different from each other at the 95% confidence interval. The difference is also economically significant. Extrapolating out from the difference in the point estimates between leased and host-owned panels suggests that on average a leased system can expect to be producing 11% more power after 10 years compared to an otherwise similar solar panel system that is host-owned.

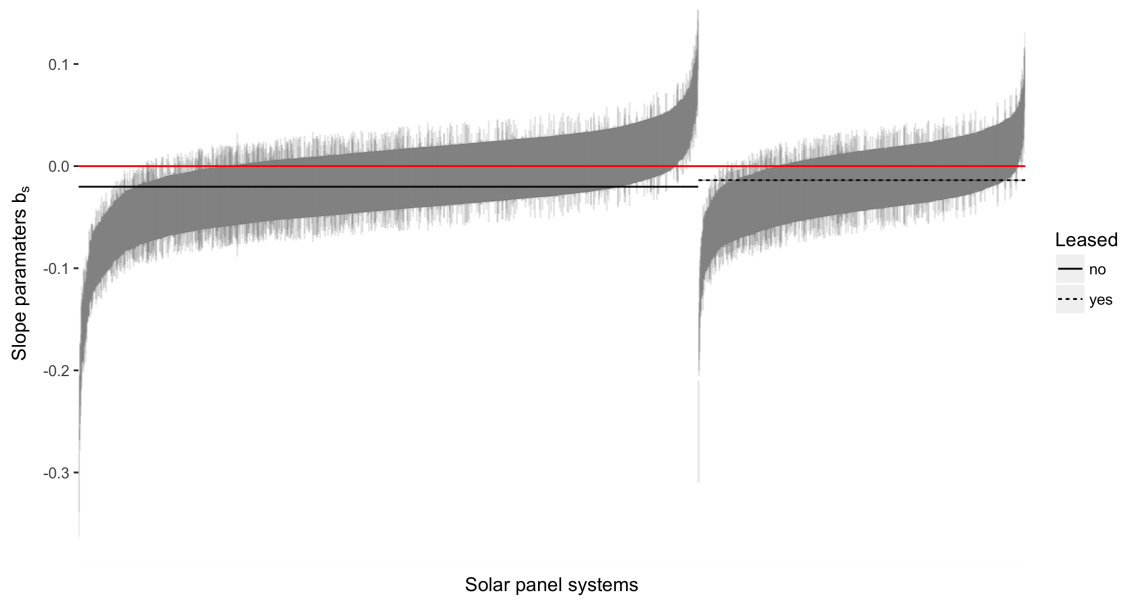


Figure 4: A visual summary of the estimation results where panels are grouped by owner type: leased (high information third-party owner) or host-owned (low information owner). The vertical lines again represent 95% confidence interval on the estimated system-level slope coefficients, which are sorted first by ownership type (host-owned on the left, leased on the right), and then from low to high of the maximum likelihood point estimate of the slope estimates. The red line represents the zero threshold. The horizontal black line on the left represents the mean of the higher level distribution on the slopes for host-owned systems,  $\mu_b[l = 0]$ , while the dotted black line on the right represents the mean of the higher level distribution on the slopes for third-party owned systems,  $\mu_b[l = 1]$

## 4 Full Bayesian hierarchical model

Up until now, I have used relatively parsimonious models with few fixed effects and only one or two hierarchies (“random effects”). Ideally, I would like to estimate a full probability model with all relevant and available covariates and hierarchies. First and foremost, so that the parameters of interests to the testable implications can be estimated simultaneously from one model. But also, as Barr et al. (2013) shows, mixed-effects models that have the maximum random effects structure justified by the model, provide the best out-of-sample prediction. Estimating mixed-effects models by maximum likelihood is efficient and straight-forward. However, convergence of the maximum likelihood routine is not guaranteed under richer models with more hierarchies and larger number of parameters to be estimated.

I wish to estimate a model where the variable for cost per-kW is included as an interaction that varies by system and module manufacturer. I also wish to have a model where both system, module manufacturer, and leasing hierarchies are estimated simultaneously. Geographic information on the location of solar panels could also conceivably be correlated with degradation of solar panels and should be accounted for in the model. I am not able to obtain convergence of the maximum likelihood routines under this rich model.

As a solution, I build a fully Bayesian hierarchical model<sup>5</sup> estimated by Markov Chain Monte Carlo (MCMC) simulation techniques.

To estimate the parameters of the model, I use the Stan Bayesian programming language and simulator (Stan Development Team, 2014), which uses Hamiltonian MCMC to estimate a joint posterior distribution of the parameters. Kruschke (2014) provides an accessible explanation of the Hamiltonian MCMC algorithm, while more detailed technical descriptions are provided by Gelman et al. (2013) and Stan Development Team (2014).

Bayesian methods making use of MCMC simulation are still emerging and this is an active field of research in both the statistical sciences and operations research fields. It is worth spending a moment to motivate their use in this case over asymptotic methods that are more common. The main reason for using the Bayesian model is the flexible treatment it allows for modeling the inherent hierarchy of the data and allowing for partial pooling in estimation.

In a hierarchical model, lower-level parameters within a certain group are themselves modelled as coming from a distribution characterised by “meta-parameters”. Thus, the estimated group-level parameters are a weighted function of both the observations within the group as well as the full set of observations in the data-set. For example, the estimated mean parameter on a group with a few, outlying observations would be pulled towards the full sample mean. This partial pooling also serves as a natural form of parameter shrinkage, which mostly eliminates the need for using corrections for multiple comparisons in inference (Gelman and Hill, 2006).

With a Bayesian multilevel model, I also have greater flexibility in specifying the model without having

---

<sup>5</sup>I purposefully change my nomenclature here and refer to the Bayesian model as a “hierarchical mode” to distinguish it from the models estimated by maximum likelihood, which I refer to as a linear mixed effects model. This distinction is mostly semantic. Both models resemble each other in structure. "hierarchical model" is however a broader term. There is no restriction to linearity. Models with non-linear or non-parametric components can also be incorporated and estimated.

to rely on assumptions like constant group-level variances and a Gaussian noise distribution. Because Bayesian simulation techniques result in an estimate of the full joint probability distribution of the model, the inference has the potential to be more informative than the typical point estimates and p-values of the standard hypothesis testing frameworks (Kruschke, 2014).

Another advantage is that the posterior distributions of the parameters have a natural and direct interpretation as probabilities, as opposed to standard hypothesis testing where the concepts of p-values and significance are often ill-defined. See Kruschke (2014) or Gelman et al. (2013) for further discussion.

Hierarchical models estimated using Bayesian techniques have some nice theoretical properties. However, the main advantages to using the Bayesian techniques here are practical. Models of nearly arbitrary richness can be estimated, with computational time and available data being the main constraints. Inference is also well-defined and exact.<sup>6</sup>

The model is described by the equations in 7. The response variable is production data that has been de-scaled as described earlier. This response variable is given a normal likelihood with mean,  $\hat{y}$  and standard deviation,  $\sigma$ . At the observation level, the fitted values,  $\hat{y}$ , are modelled as a simple linear regression, with an intercept  $a_s$ , and a slope  $b_s$ , which are both allowed to vary by system,  $s$ . In addition a vector of month dummies,  $\mathbf{month}_c$ , is included in order to control for seasonality. In contrast to the earlier estimation, I allow the vector of coefficients on the month dummies to vary by county, indexed by  $c$ . California is a large, geographically diverse state, and the seasonal profile of solar generation is bound to vary across counties.

At the system level, the intercept terms,  $a_s$  are given a common higher level normal distribution with mean  $\mu^a$  and standard deviation  $\sigma^a$ .

The slope terms,  $b_s$  are also modelled as normal distributions with mean terms  $\mu_s^b$  and standard deviations  $\sigma_s^b$ , which are both allowed to vary by system. I decompose the variation of  $\mu_s^b$  by groupings and covariates of interest. I account for variation by lease,  $\mu_l^{lease}$ , host sector  $\mu_c^{sect}$ , and manufacturer  $\mu_m^{manuf}$ . In addition, three continuous variables are included in the system-level equation: the year of first production,  $first\_prod\_year_s$ , capacity,  $capacity_s$  and the system cost per kilowatt,  $cost_s$  with corresponding coefficients  $\beta^{fy}$ ,  $\beta^{size}$ , and  $\beta^{cost}$ . The coefficients in this equation can be interpreted as interaction effects with the slope terms as in a traditional regression. It can also be helpful to think of it as a two-stage regression, where the  $b_s$  terms are first estimated and interpreted as a proxy for the quality of a solar power system. The estimated  $\hat{b}_s$  terms are in turn regressed on the explanatory variables. In reality, estimation is done simultaneously and the  $b_s$  parameters are appropriately regularized by the higher-level distributions.

---

<sup>6</sup>Exact in the sense that given a correctly specified probability model, then an MCMC routine that is given infinite time to sample, will in theory converge to the correct posterior probability distribution. Maximum likelihood is approximate in the sense that inference is based on asymptotic theory and assumptions of normality.



Symbol	Definition
$prod_i$	Production, descaled, from a solar power system at a given month (observation i).
$months\_operation_{st}$	Number of months of operation of solar system s at month t
$\mathbf{month_c}$	Vector of month dummy variables that can vary by county, c
$first\_prod\_year_s$	Year of first production from solar system, s
$log\_csi\_rating_s$	Size of solar power system, as measured by log production capacity, descaled, of solar power system, s
$log\_cost_s$	Log cost, descaled, of solar power system, s
$\hat{y}_{st}$	Mean of the response variable, production for system s at time t
$\sigma$	Standard deviation of the response variable, production
$a_s$	Intercept parameter for a system, s
$b_s$	Slope parameter for a system, s
$\mu^a$	Mean of the distribution of system intercept terms, $a_s$
$\mu_s^b$	Mean of the distribution of system slope terms, $b_s$ .
$\sigma^a$	Variance of the distribution of system intercept terms, $a_s$
$\sigma_s^b$	Variance of the distribution of the system slope terms, $b_s$ .
	Can vary by system, s
$\mu^{month}$	Mean of the distribution of the county-varying month indicators
$\sigma^{month}$	Variance of the distribution of the county-varying month indicators
$x_0$	Mean of the higher level (meta) distribution of the slope standard deviations, $\sigma_s^b$
$\gamma$	Variance of the higher level (meta) distribution of the slope standard deviations, $\sigma_s^b$
$\mu_l^{lease}$	Mean parameters for distribution of slopes of leased and host-owned groups, l
$\mu_e^{sector}$	Mean parameters for distribution of slopes of host sector groups, e.
$\mu_m^{manuf}$	Mean parameters for distribution of slopes of panel manufacturer groups, m.
$\beta^{fy}$	Parameter on variable for first production year, $first\_prod\_year$
$\beta^{size}$	Parameter on variable for size of solar power system, $log\_csi\_rating$
$\beta^{cost}$	Parameter on variable for size of solar power system, $log\_csi\_rating$

Table 5: Definitions from the full Bayesian model presented in equation 7.

$$\begin{aligned}
prod_i &\sim N(\hat{y}_{st}, \sigma) \\
\hat{y}_{st} &= a_s + b_s months\_operation_{st} + \mathbf{month_c} \\
a_s &\sim N(\mu^a, \sigma^a) \\
b_s &\sim N(\mu_s^b, \sigma_s^b) \\
\mathbf{month_c} &\sim N(\mu^{month}, \sigma^{month}) \\
\sigma_s^b &\sim \text{half-Cauchy}(x_0, \gamma) \\
\mu_s^b &= \mu_l^{lease} + \mu_e^{sect} + \mu_m^{manuf} \\
&\quad + \beta^{fy} first\_prod\_year_s + \beta^{size} log\_csi\_rating_s + \beta^{cost} log\_cost_s
\end{aligned} \tag{7}$$

The standard deviation parameters and higher level terms are all given Cauchy prior distributions as summarized in the equations in 8. The Cauchy distribution, which is a T-distribution with 1 degree of freedom, has wide tails and thus allows for the occasional outlying coefficient in a lower level regression, while still provides some regularization towards zero (Gelman et al., 2008). The Cauchy(0,1) prior for location parameters of de-scaled variables is what is referred to as a weakly-informative prior. As opposed to a completely flat prior, it imposes some weak regularization on the parameters in order to avoid over-fitting the data and improve predictive performance. At the same time, the prior does not impose any strong prior knowledge. The half-Cauchy(0,5) prior for scale parameters is in practice coded as a Cauchy(0,5) distribution restricted to a positive support

$$\begin{aligned}
\mu_l^{lease} &\sim \text{Cauchy}(0, 1) \\
\mu_s^{sect} &\sim \text{Cauchy}(0, 1) \\
\mu_m^{manuf} &\sim \text{Cauchy}(0, 1) \\
\mu^a &\sim \text{Cauchy}(0, 1) \\
mon_c &\sim \text{Cauchy}(0, 1) \\
\beta^{fy} &\sim \text{Cauchy}(0, 1) \\
\beta^{size} &\sim \text{Cauchy}(0, 1) \\
\beta_l^{cost} &\sim \text{Cauchy}(0, 1) \\
x_0 &\sim \text{Cauchy}(0, 1) \\
\sigma &\sim \text{half-Cauchy}(0, 5) \\
\sigma^a &\sim \text{half-Cauchy}(0, 5) \\
\sigma^{mon} &\sim \text{half-Cauchy}(0, 5) \\
\gamma &\sim \text{half-Cauchy}(0, 5)
\end{aligned} \tag{8}$$

In the earlier section, we noticed that some of the system slope parameters were estimated to be positive. There are intuitive and plausible reasons for observing higher average production sometime after the initial installation. The placement of solar panels could have been optimized after the fact. Malfunctioning components could also have been replaced. However, when we wish to interpret the  $b_s$  parameters as

I	<p><b>Likelihood of the Response Data</b> Production from system, s, in period t</p> $prod_i \sim N(\hat{y}_i, \sigma)$
II	<p><b>Decomposition of Fitted Data</b> Intercept and slope that vary by system plus month-of-year dummies that control for seasonality and vary by county</p> $\hat{y}_i = a_s + b_s months\_operation_i + month_c$
III	<p><b>System level parameters and decomposition of slope term</b> Slope terms modelled with location parameter that is a linear function of system-level variables</p> $a_s \sim N(\mu^a, \sigma^a)$ $b_s \sim N(\mu_s^b, \sigma_s^b)$ $month_c \sim N(\mu^{month}, \sigma^{month})$ $\mu_s^b = \mu_l^{lease} + \mu_s^{sect} + \mu_m^{manuf} + \beta^{fy} first\_prod\_year_s + \beta^{size} csi\_rating_s + \beta^{cost} cost_s$
IV	<p><b>Meta-parameters</b> Regularizing higher level distributions</p> $\mu^k \sim Cauchy(0,1) \quad k \in \{lease, sect, manuf, a, mon\}$ $\beta^i \sim Cauchy(0,1) \quad i \in \{fy, size, cost\}$ $\sigma_s^b \sim half - Cauchy(x_0, \gamma)$ $\sigma^f \sim half - Cauchy(0,5) \quad f \in \{a, \mu, \beta\}$

Figure 5: An overview of the structure of the model. Starting from the top in panel I, the monthly production data is modelled with a Normally distributed likelihood. Moving down to II, the fitted values,  $\hat{y}_i$ , are in turn decomposed into system-varying slope and intercept terms, as well as month-dummies that control for seasonality. Further down to III, the slope terms,  $b_s$  are modelled as Normal distributions with varying mean parameters,  $\mu_s^b$  that decompose further into terms representing variables of interest and control variables. Finally, in IV, all parameters are given higher-level meta-parameters that serve to regularize (or "shrink") the lower-level parameter estimates.

degradation over time, we should ideally discount positive values. Several possibilities present themselves. I could discard the systems with positive slope parameters from the data. But discarding data is generally not desirable. Alternatively, I could keep in all data, allowing for the slope terms to be positive. One of the advantages of the hierarchical model with weakly regularizing priors is that it allows for outlying parameter estimates, without unduly affecting the overall results. A third solution is to include all the data but constrain the model by specifying priors that place zero probability on positive slope parameter values. In particular the  $b_s$  parameters would be given Cauchy distributions with support between 0 and negative infinity. This has the desired effect of constraining slope parameters to be negative by imposing strong prior knowledge. The downside is that such strong priors can sometimes lead to unintended effects on the posterior distribution and faulty inference. In the Bayesian literature, such hard constraints are generally not recommended. In the proceeding analysis I therefor present results from both constrained and unconstrained models.

#### 4.1 Computation and Results from Bayesian hierarchical model

The model is estimated using the Stan probabilistic programming language and sampler, which implements Hamiltonian MCMC and a No U-turn sampler (NUTS) (Stan Development Team, 2014). The sampler was run with four chains and 1000 iterations. Gelman-Rubin convergence statistics ( $\hat{R}$ ) were used to check convergence in the model. 2000 samples are taken from the simulated posterior distribution, and model results for the marginal distributions of the parameters of interest are presented in the form of kernel densities of the samples.

- Recall that the testable implications from the theory of asymmetric information and quality were that
- 1.) average quality should be higher in the panels owned by the high-information types (leased panels).
  - 2.) Price should be more highly correlated with quality in the panels owned by high-information types.

The parameters of interest in our model are then the higher-level parameters  $\mu_i^{lease}$  and  $\beta_i^{cost}$ . The first represents the mean value of the system-level slope terms,  $b_s$ , grouped by whether the systems were leased or host-owned. The latter is a measure of the correlation between the cost of the system and the system slope terms. The  $\beta_i^{cost}$  parameter is also allowed to vary by leased or host-owned. Importantly, the  $\beta_i$  parameters are estimated conditional on the installation date of the solar panels; thus they do not reflect changes in price over time. This could lead to a heavy bias in the results if, as seems plausible, both prices and quality changed over time.

For the testable implications, the inference we are interested in is the contrasts,  $\mu_{lease,yes} - \mu_{lease,no}$  and  $\beta_{leased} - \beta_{owned}$ . Kernel densities representing the marginal distributions of these contrasts are presented in figure 6. The top panel shows the distribution of the contrast  $\mu_{lease,yes} - \mu_{lease,no}$ . Approximately 99% of the probability mass is located in the positive range. This reinforces the finding from the mixed effects model that leased panels do appear to have substantially higher quality, as the theory of asymmetric information of quality would suggest.

The evidence for the second testable implication is less conclusive. The model estimates that there is approximately a 85% probability that there is a higher correlation between quality and cost for the high

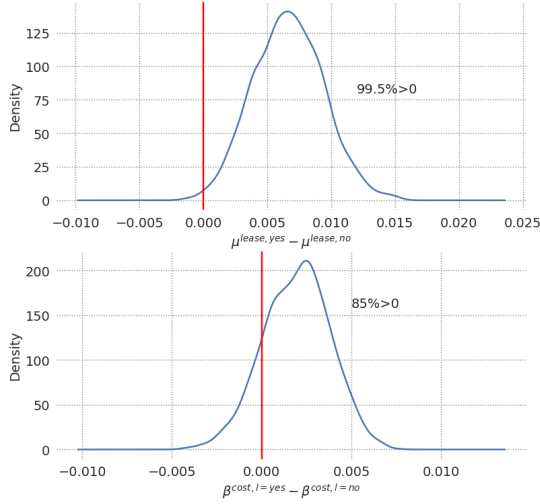


Figure 6: Marginal posterior distributions for the contrasts of interest. The top panel shows that the distributions of the slope parameters,  $b_s$  between leased and host-owned systems as defined by the contrast,  $\mu_{lease,yes} - \mu_{lease,no}$  have approximately 99% probability of being positively different from each other. The lower panel shows the marginal posterior distribution of the contrast between the  $\beta_l^{cost}$  parameter for leased and host-owned solar systems. The results indicate that there is an approximately 85% probability that costs are more highly correlated with quality in leased systems compared to those that are host-owned.

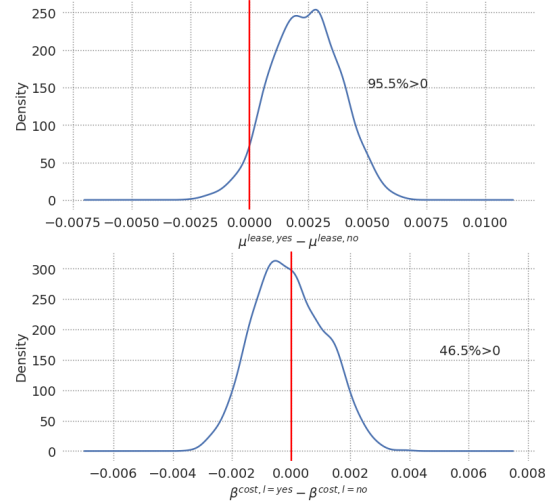


Figure 7: Marginal posterior distributions for the contrasts when the  $b_s$  parameters and corresponding higher-level mean parameters are constrained to have a maximum of zero. The contrast,  $\mu_{lease,yes} - \mu_{lease,no}$  is pulled somewhat towards zero, but continues to have approximately 95% percent of its probability mass in the positive range. The contrast  $\beta_{lease,yes} - \beta_{lease,no}$  is also pulled towards zero, and the distribution of the contrast is closely centered around zero.

information type owner of leased panels compared to the low information type owner of host-owned panels.

Figure 7 show the corresponding distributions under the constrained model.  $\mu_{lease,yes} - \mu_{lease,no}$  is pulled somewhat towards zero in magnitude, but about 95% of the probability mass is still in the positive range. The distribution of  $\beta_{leased} - \beta_{owned}$  is also pulled towards zero, and now is centered close to zero.

Interpretation of the  $\beta_l^{cost}$  parameters must be done with some care. A direct comparison of costs between leased and host-owned solar power systems is somewhat complicated. The prices of host-owned panels are market prices: Simply the purchase price that the owner of the solar panel system paid to the contractor. On the other hand, the price of leased systems is a reported price from the contractors consisting of the component and installation costs plus an estimated mark-up. If these costs were reported correctly, then the theory should still hold. The reported cost should reflect the ability of the high-information owners to judge quality of the panels that they purchase. In practice, the reliability of reported cost is a source of added uncertainty and should be taken into consideration when interpreting the results.

However, if we assume that misreporting of the true costs will tend to lead to less variation in reported costs, then that should have the effect of attenuating the magnitude of the contrast,  $\beta_{cost,l=yes} - \beta_{cost,l=no}$ . In that case, the estimated contrast can be seen as being a conservative estimate.

Summary statistics of the distributions of higher level parameters are reported in table 6 and table 7 for the constrained model. The sector groupings,  $\mu^{sect}$  can be of interest in a discussion of the role of

	mean	std	min	2.5%	25%	50%	75%	97.5%	max
$\mu^a$	0.007	0.052	-0.116	-0.094	-0.026	0.012	0.040	0.095	0.124
$\sigma^a$	0.907	0.012	0.872	0.885	0.899	0.907	0.915	0.934	0.952
$\sigma^{cm}$	0.299	0.009	0.268	0.280	0.293	0.298	0.305	0.315	0.331
$\sigma$	0.321	0.001	0.319	0.320	0.320	0.321	0.321	0.322	0.323
$\mu^{lease,no}$	-0.005	0.370	-1.008	-0.824	-0.239	-0.028	0.271	0.705	0.936
$\mu^{lease,yes}$	0.001	0.370	-1.000	-0.814	-0.232	-0.022	0.278	0.716	0.943
$\mu^{sect,com}$	-0.048	0.348	-0.995	-0.752	-0.300	-0.026	0.163	0.707	0.887
$\mu^{sect,res}$	-0.049	0.348	-0.998	-0.753	-0.302	-0.027	0.162	0.708	0.885
$\mu^{sect,gov}$	-0.045	0.348	-0.990	-0.748	-0.298	-0.024	0.168	0.716	0.889
$\mu^{sect,npr}$	-0.049	0.348	-0.999	-0.756	-0.300	-0.025	0.163	0.706	0.893
$\beta^f_y$	-0.006	0.002	-0.013	-0.010	-0.007	-0.006	-0.005	-0.003	-0.001
$\beta^{cost}_{own}$	-0.001	0.001	-0.005	-0.003	-0.002	-0.001	0.000	0.002	0.004
$\beta^{cost}_{lease}$	0.001	0.002	-0.005	-0.002	0.000	0.001	0.002	0.004	0.006
$\beta^{size}$	-0.021	0.001	-0.025	-0.024	-0.022	-0.021	-0.020	-0.018	-0.016

Table 6: Summary statistics of the estimated posterior distributions of higher level parameters

	mean	std	min	2.5%	25%	50%	75%	97.5%	max
$\mu^a$	-0.037	0.035	-0.129	-0.104	-0.065	-0.031	-0.010	0.021	0.045
$\sigma^a$	0.907	0.012	0.868	0.884	0.899	0.906	0.915	0.930	0.946
$\sigma^{cm}$	0.298	0.010	0.269	0.280	0.291	0.297	0.304	0.318	0.333
$\sigma$	0.328	0.001	0.326	0.327	0.328	0.328	0.329	0.329	0.330
$\mu^{lease,no}$	0.163	0.221	-0.589	-0.404	0.060	0.181	0.291	0.578	0.695
$\mu^{lease,yes}$	0.165	0.221	-0.585	-0.399	0.060	0.183	0.294	0.579	0.700
$\mu^{sect,com}$	-0.240	0.216	-0.788	-0.629	-0.394	-0.247	-0.123	0.246	0.495
$\mu^{sect,res}$	-0.242	0.216	-0.789	-0.628	-0.395	-0.250	-0.122	0.244	0.493
$\mu^{sect,gov}$	-0.239	0.216	-0.782	-0.625	-0.392	-0.245	-0.120	0.250	0.501
$\mu^{sect,npr}$	-0.240	0.216	-0.786	-0.629	-0.396	-0.249	-0.121	0.250	0.498
$\beta^f_y$	-0.003	0.001	-0.006	-0.004	-0.003	-0.003	-0.002	-0.001	0.000
$\beta^{cost}_{own}$	-0.000	0.001	-0.002	-0.002	-0.001	-0.000	0.000	0.001	0.002
$\beta^{cost}_{lease}$	-0.000	0.001	-0.003	-0.002	-0.001	-0.000	0.000	0.002	0.003
$\beta^{size}$	-0.007	0.001	-0.011	-0.009	-0.008	-0.007	-0.006	-0.005	-0.004

Table 7: Summary statistics of the estimated posterior distributions of higher level parameters when  $b_s$  parameters are constrained to have a maximum of zero.

asymmetric information. Figure 8 shows the distributions of the contrasts relative to the residential host sector. The top panel represents the contrast with commercial sector hosts, the middle with government sector host and the bottom panel with non-profit sector host. The contrasts for commercial and non-profit hosts relative to residential have distributions centered close to zero. There is little evidence that quality varies significantly. The distribution for the contrast between government and residential solar panels has about 90 percent of its probability mass in the positive range. We can interpret this to say that there is a 90% probability that solar panels systems with governmental hosts display on average higher quality than similar solar panels systems with residential hosts. Results of the constrained model are shown in figure 9. The results are similar.

While it is not clear why solar panel systems with governmental hosts should display higher quality, an explanation consistent with the theory of asymmetric information of quality is that governmental host may be better informed. Local or state governments are likely repeat buyers of solar panel systems and have acquired knowledge about quality.

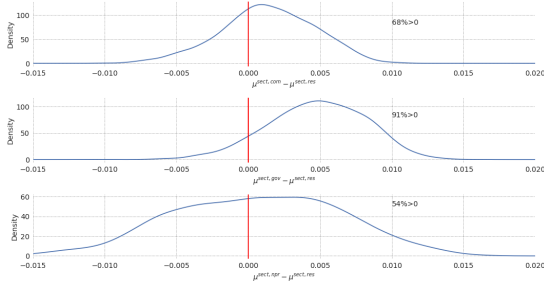


Figure 8: Distribution of the sector contrasts relative to residential sector. Installations for commercial and non-profit hosts do not appear to have on average significant differences in quality compared to residential installations. There is estimated approximately 90 percent probability that governmental installations have a higher quality than residential installations

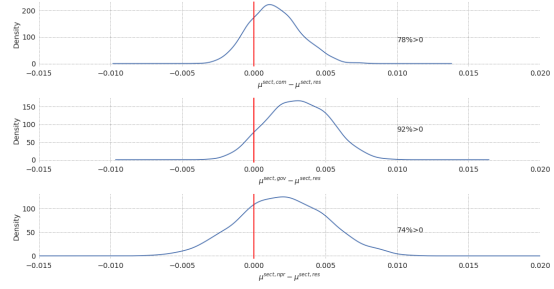


Figure 9: Distribution of the sector contrasts relative to residential sector when all higher level mean parameters are constrained to have a maximum of zero. Results are not materially changed.

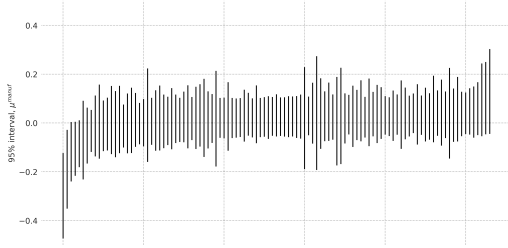


Figure 10: Plot of the the parameters  $\mu^{manuf}$ , representing the distributions of the mean parameter for manufacturer groupings.

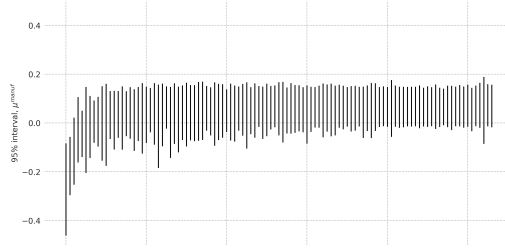


Figure 11: Plot of the the parameters  $\mu^{manuf}$ , when all higher level mean parameters have a maximum of zero.

Summary statistics for the parameters on the manufacturer group means as well as the lower-level intercept and slope parameters  $a_s$  and  $b_s$  are impractical to present in a table. Visual summaries are presented in figures 10, 12 and 14. Figure 10 shows line plots of the manufacturer distributions, with the lines representing 95 percent intervals of the probability distribution. Figure 11 shows results from the constrained model. Mirroring earlier results, the plot shows most manufacturers with roughly equivalent quality, with a handful with meaningfully lower quality. Figures 12 and 14 show 95% intervals of the distribution of the  $a_s$  and  $b_s$  parameters. The intercept terms are shown to have considerable variation between systems, as would be expected, but are individually estimated precisely. Figures 13 and 15 show results from the constrained model.

The slope parameters,  $b_s$  are modelled explicitly and are also allowed to have varying variance parameters. This is evident in the plot. A similar pattern to the higher-level manufacturer group means is apparent: modest differences in quality are seen between most systems, with the exception of a lower tail of systems with particularly poor quality.

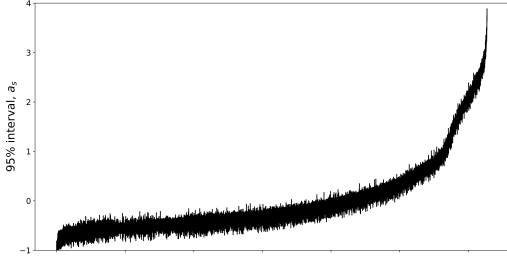


Figure 12: 95% interval of the distributions of the system level intercept parameters,  $a_s$ .

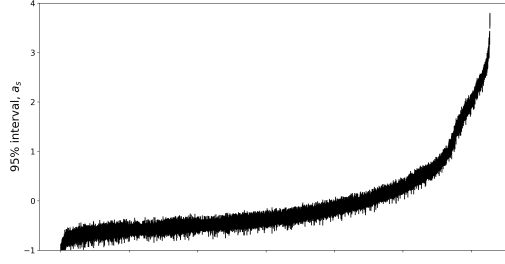


Figure 13: 95% interval of the distributions of the system level intercept parameters,  $a_s$  when higher level mean parameters are constrained to have a maximum of zero.

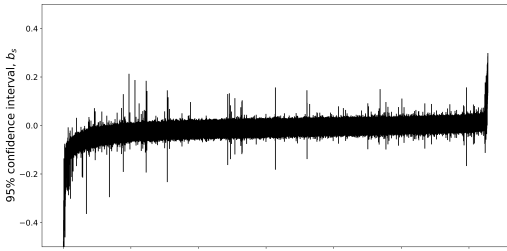


Figure 14: 95% interval of the distributions of the system level slope parameter,  $b_s$

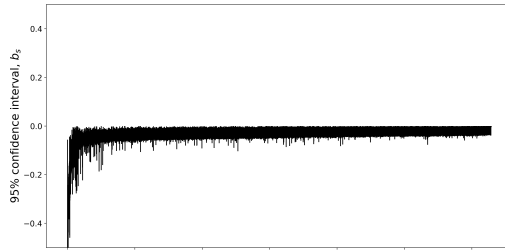


Figure 15: 95% interval of the distributions of the system level slope parameter,  $b_s$  when distributions are constrained to be at or below 0.

## 5 Discussion and Conclusion

Because of the dramatic price declines in solar panels and the competitive pressure that western producers have faced, it has become common to hear in the industry that solar panels have become commodities. But in finance and economics, commodity has a specific meaning. Commodities of a certain grade need to be fungible: goods from different producers are interchangeable.

If solar panels truly have become commodities, it would have meaningful and substantial implications for the solar power industry. For example, financial instruments could arise providing developers of solar power plants or solar panel retailers the ability to hedge price uncertainty in panels and guarantee supplies for projects with long lead times. Particularly relevant to the current environment would be trade policy. Trade restrictions on panels from specific countries, like the current US tariffs on Chinese produced panels, would presumably be less effective. Fungible panels from China could easily be sold elsewhere, while similar panels from other countries supply the US market.

Yet the findings from this article strongly suggest that solar panels are not commodities. In particular, I find that the salient property of quality can vary substantially between producers. Because solar power is a distributed technology, where assets can be owned by individual homeowners, businesses and other organizations that do not have the resources to judge quality, the issue of asymmetric information may be present in the market.

From a descriptive understanding of the market, I identify a high information type owner (third-



party owners) and low information type owner (host-owners) of solar power assets. Then, from the basic theory of asymmetric information of quality, I explore two testable implications. I find that, as the theory would predict, high information owners are able to attain higher-quality panels. With about 85 percent probability, I also find that the cost of solar panel systems owned by high-information types is more highly correlated with quality, as theory would also suggest. It is important to note that while the empirical analysis is consistent with the results predicted by the theory of asymmetric information, asymmetric information is not perfectly identified. The results could plausibly be driven by other factors. For example, the different ownership models could induce a moral-hazard type behavior when it comes to maintenance of the solar panels, which in turn could effect production over time.

The market for solar panels and solar panel systems is however still young and dynamic. One of the advantages of the linear mixed-effects models and the hierarchical models is that we can also get an idea of the degree of differentiation in quality between different manufacturers. About 20% of producers accounted for the worst quality panels, while the majority of panels had similar quality. If low-quality producers either are forced to exit the market or improve their quality to match standards in the industry, then it is possible that solar panels could indeed become commodities in the future.

However, optimism should be tempered by both theory and experience from other industries. Markets that are subject to informational problems have a tendency to experience stasis. As Akerlof (1970) notes, poor quality producers will tend to push out high quality producers in the presence of asymmetric information. In a long-term study of the used car market in Switzerland, Emons and Sheldon (2009) find evidence of sustained quality uncertainty among buyers. Evidence of long-term quality uncertainty is also found in the market for business aircraft (Gilligan, 2004). The market for solar panels may long be characterised by producers differentiated by quality. In consequence, the buyers of solar panels will need to take steps to ensure high quality such as paying a premium for panels from a reputed producer or making use of engineering firms specialising in certifying panel quality.

The use of Bayesian MCMC simulation techniques in this article also provides an illustrative example of the use and usefulness of a powerful emerging tool-set. The Bayesian frameworks and recently developed software such as Stan provide major practical and theoretical advantages over many traditional statistical methods used in operations research and decision science. These methods and tools are also increasingly accessible to practitioners. Kruschke (2014), McElreath (2015), Gelman et al. (2013) and Stan Development Team (2014) provide accessible and thorough introductions and guides.

This article should also be considered as part of a broader literature on the special characteristics of new distributed energy generation technologies. The energy investment behavior of home owners, farmers, and small cooperatives are bound to be substantially different than those of large, specialized energy companies that have traditionally done most of the investment in electricity generation. Understanding the quickly evolving energy and power industry requires taking into account informational and behavioral factors, and this is a rich field for further research in the field of operations research.

## References

- Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.
- Bates, D., Machler, M., Walker, S. C., and Bolker, B. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis 2nd (second) edition Text Only*. Springer.
- Bhattacharya, A., Wilson, S. P., and Soyer, R. (2019). A Bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research*, 274(3):1112–1124.
- Bodnar, T., Mazur, S., and Okhrin, Y. (2017). Bayesian estimation of the global minimum variance portfolio. *European Journal of Operational Research*, 256(1):292–307.
- Bollinger, B. and Gillingham, K. (2012). Peer Effects in the Diffusion of Solar Photovoltaic Panels. *Marketing Science*, 31(6):900–912.
- Caulkins, J. P., Feichtinger, G., Grass, D., Hartl, R. F., Kort, P. M., and Seidl, A. (2017). Interaction of pricing, advertising and experience quality: A dynamic analysis. *European Journal of Operational Research*, 256(3):877–885.
- Chan, Y.-S. and Leland, H. (1982). Prices and Qualities in Markets with Costly Information. *The Review of Economic Studies*, 49(4):499–516.
- Cronon, W. (1992). *Nature's Metropolis: Chicago and the Great West*. W. W. Norton & Company, New York, new ed edition edition.
- Dastrup, S. R., Graff Zivin, J., Costa, D. L., and Kahn, M. E. (2012). Understanding the Solar Home price premium: Electricity generation and “Green” social status. *European Economic Review*, 56(5):961–973.
- Dehejia, R. H. (2003). Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs With Grouped Data. *Journal of Business & Economic Statistics*, 21(1):1–11.
- Dobbie, W. and Skiba, P. M. (2013). Information Asymmetries in Consumer Credit Markets: Evidence from Payday Lending. *American Economic Journal: Applied Economics*, 5(4):256–282.
- Emons, W. and Sheldon, G. (2009). The market for used cars: new evidence of the lemons phenomenon. *Applied Economics*, 41(22):2867–2885.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, Boca Raton, 3 edition edition.

- Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge ; New York, 1 edition edition.
- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.
- Gensler, S., Hinz, O., Skiera, B., and Theysohn, S. (2012). Willingness-to-pay estimation with choice-based conjoint analysis: Addressing extreme response behavior with individually adapted designs. *European Journal of Operational Research*, 219(2):368–378.
- Gilligan, T. W. (2004). Lemons and Leases in the Used Business Aircraft Market. *Journal of Political Economy*, 112(5):1157–1180.
- Gordon, D. V., Hannesson, R., and Kerr, W. A. (1999). What is a Commodity? An Empirical Definition Using Time Series Econometrics. *Journal of International Food & Agribusiness Marketing*, 10(2):1–29.
- Healy, P. M. and Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31(1–3):405–440.
- Homan, M. D. and Gelman, A. (2014). The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.
- Jones, C. S. and Shanken, J. (2005). Mutual fund performance with learning across funds. *Journal of Financial Economics*, 78(3):507–552.
- Karniouchina, E. V., Moore, W. L., van der Rhee, B., and Verma, R. (2009). Issues in the use of ratings-based versus choice-based conjoint analysis in operations management research. *European Journal of Operational Research*, 197(1):340–348.
- Kirmani, A. and Rao, A. R. (2000). No Pain, No Gain: A Critical Review of the Literature on Signaling Unobservable Product Quality. *Journal of Marketing*, 64(2):66–79.
- Korteweg, A. G. and Sorensen, M. (2017). Skill and luck in private equity performance. *Journal of Financial Economics*, 124(3):535–562.
- Kruschke, J. K. (2014). *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*. Academic Press, Boston, 2 edition edition.
- Le Cadre, H., Mezghani, I., and Papavasiliou, A. (2019). A game-theoretic analysis of transmission-distribution system operator coordination. *European Journal of Operational Research*, 274(1):317–339.
- Liu, F., Hua, Z., and Lim, A. (2015). Identifying future defaulters: A hierarchical Bayesian method. *European Journal of Operational Research*, 241(1):202–211.
- MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK ; New York, 1 edition edition.

- Mauritzen, J. (2017). Cost, Contractors and Scale: An Empirical Analysis of the California Solar Market. *The Energy Journal*, Volume 38(Number 6).
- McElreath, R. (2015). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Boca Raton, 1 edition edition.
- Meager, R. (2018). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, Forthcoming.
- Menger, C., Klein, P. G., Hayek, F. A. v., Dingwall, J., and Hoselitz, B. F. (2007). *Principles of economics*. Ludwig von Mises Institute, Auburn, Ala. OCLC: 123178515.
- Michaely, R. and Shaw, W. H. (1994). The Pricing of Initial Public Offerings: Tests of Adverse-Selection and Signaling Theories. *Review of Financial Studies*, 7(2):279–319.
- Petersen, M. A. and Rajan, R. G. (1994). The Benefits of Lending Relationships: Evidence from Small Business Data. *The Journal of Finance*, 49(1):3–37.
- Rufo, M. J., Perez, C. J., and Martin, J. (2010). Merging experts’ opinions: A Bayesian hierarchical model with mixture of prior distributions. *European Journal of Operational Research*, 207(1):284–289.
- Shapiro, C. (1982). Consumer Information, Product Quality, and Seller Reputation. *The Bell Journal of Economics*, 13(1):20–35.
- Soyer, R. and Tanyeri, K. (2006). Bayesian portfolio selection with multi-variate random variance models. *European Journal of Operational Research*, 171(3):977–990.
- Stan Development Team (2014). *Stan Modeling Language Users Guide and Reference Manual, Version 2.2*.
- Szwed, P., Dorp, J. R. v., Merrick, J. R. W., Mazzuchi, T. A., and Singh, A. (2006). A Bayesian paired comparison approach for relative accident probability assessment with covariate information. *European Journal of Operational Research*, 169(1):157–177.
- Tirole, J. (1988). *The Theory of Industrial Organization*. The MIT Press, Cambridge, Mass, 1st edition edition.
- Utkin, L. V. (2006). A method for processing the unreliable expert judgments about parameters of probability distributions. *European Journal of Operational Research*, 175(1):385–398.
- Velik, R. and Nicolay, P. (2016). Energy management in storage-augmented, grid-connected prosumer buildings and neighborhoods using a modified simulated annealing optimization. *Computers & Operations Research*, 66:248–257.