Strategic Management Journal
Strat. Mgmt. J., **37**: 86–97 (2016)
Published online EarlyView 8 December 2015 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2450
Received 31 August 2013; *Final revision received 24 November 2014*

THE USE OF SPARSE INVERSE COVARIANCE ESTIMATION FOR RELATIONSHIP DETECTION AND HYPOTHESIS GENERATION IN STRATEGIC MANAGEMENT

MEI LI,[1] YING LIN,[2] SHUAI HUANG,[2] and CRAIG CROSSLAND[1]*
[1] *Department of Management, University of Notre Dame, Notre Dame, Indiana, U.S.A.*
[2] *Department of Industrial and Systems Engineering, University of Washington, Seattle, Washington, U.S.A.*

Research summary: *This paper uses Sparse Inverse Covariance Estimation (SICE) to advance strategic management research, focusing on an application of exploratory SICE techniques to generate novel, testable hypotheses. We demonstrate how SICE can identify intrinsic relationships among variables, especially within large, high-dimension–low-observation datasets. We also discuss the strengths and limitations of SICE, as well as the most appropriate uses of these techniques. We conclude with a detailed illustration of SICE analysis using the High Performance Manufacturing dataset.*
Managerial summary: *Most academic research in strategic management is based on formal hypotheses, which are testable statements of the relationships linking two or more variables. Researchers currently use several approaches to generate hypotheses, but all have different shortcomings. In our study, we demonstrate how researchers can use a set of quantitative techniques known as Sparse Inverse Covariance Estimation (SICE) to generate new hypotheses from large datasets.*

INTRODUCTION

The process of generating new scientific knowledge is complex and uncertain. In most empirical research underpinned by assumptions of ontological and epistemological realism (Kilduff, Mehra, and Dunn, 2011), a crucial step in this process is the development of one or more hypotheses—falsifiable statements concerning the relationships among a set of variables (Popper, 1959). Researchers tend to approach this task in a number of different ways. Most simply, one can engage in deductive puzzle solving within an established paradigm (Kuhn, 1996). In this approach, the researcher begins with an existing theory and deduces novel hypotheses based on hitherto untested implications of the theory. Alternatively, researchers can identify new research opportunities through induction (Kilduff, 2006). In this approach, the researcher begins with specific real-world observations or phenomena and then postulates a more general relationship (Lakatos, 1970). Many qualitative techniques, such as grounded theory (e.g., Gioia and Chittipeddi, 1991), represent inductive approaches to building theory and/or hypotheses from observational data, often within the context of a small number of firms.

However, to this point, little work provides guidance as to how strategic management researchers

Keywords: sparse inverse covariance estimation; Gaussian graphical models; hypothesis generation; relationship detection in high-dimension–low-observation data; exploratory studies
*Correspondence to: Craig Crossland, 328 Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556 USA.
E-mail: craigcrossland@gmail.com

might systematically use the increasing amounts of quantitative firm-level data at their disposal to generate new hypotheses, while simultaneously avoiding the statistical and ethical pitfalls of "searching for asterisks" (Bettis, 2012). Although researchers have become increasingly sophisticated in the use of analytical techniques to evaluate causal hypotheses (e.g., Bascle, 2008; Semadeni, Withers, and Certo, 2014), much less consideration has been devoted to the issue of generating suitable hypotheses in the first place. In other words, strategy researchers are getting better answers to their questions, but it is not clear that these are always the best questions to ask (Mahoney and McGahan, 2007).

In our study we address this challenge by showing how researchers can use Sparse Inverse Covariance Estimation (SICE), a set of exploratory techniques drawn from the biological and physical sciences, to expand the range of original questions available to them. We discuss how SICE can be used to identify the fundamental (intrinsic) bivariate relationships among variables within a dataset, and thus assist researchers to generate novel, testable hypotheses. SICE is especially useful in situations where the dataset is very large and the number of potential variables of interest is substantially higher than the number of units being considered (i.e., $p > n$).

In the following sections, we provide a description of SICE and its uses, followed by a detailed example of applying SICE techniques to the High Performance Manufacturing dataset ($p = 1697$; $n = 197$). Also, see Online Supplemental Material (Appendix S1) for a technical description of SICE and a mathematical proof, along with the MATLAB programming codes used in this study.

SPARSE INVERSE COVARIANCE ESTIMATION

SICE techniques are used to uncover the intrinsic connectivity among a particular network of variables (Friedman, Hastie, and Tibshirani, 2008, 2010; Huang *et al.*, 2010; Jones *et al.*, 2012; Weiss and Freeman, 2001). Originally developed by Dempster (1972), SICE techniques have been used in a range of populations, including brain regions (Huang *et al.*, 2010; Valdes-Sosa *et al.*, 2005), speech patterns (Zhang and Fung, 2013), equities (Fan, Lv, and Qi, 2011), genes (Dobra *et al.*, 2004; Toh and Horimoto, 2002), people (Ahmed and Xing, 2009; Myers and Leskovec, 2010), and organizations (Kim *et al.*, 2013).

A simple example illustrates the aims and purpose of SICE analyses. Suppose we have a network of three variables: a firm's overall vision for innovation, a firm's policies relating to the hiring and promotion of innovative employees, and a firm's incentives supporting innovation. A firm's innovation vision is, separately, correlated with its selection practices and its incentive structure. However, in most samples, the observed relationships are likely to be reflected with an additional (spurious) relationship between selection and incentive structure. In large datasets, where the number of variables can be substantial, spurious relationships such as these make it challenging to determine the most important core relationships. SICE helps to eliminate the redundancies in relationship mapping and thereby uncover intrinsic patterns among variables.

The working logic for SICE analysis is as follows. First, an empirical covariance matrix is computed based on the observed relationships among all variables within a dataset. Next, conditioned on the empirical covariance matrix, a likelihood function of the inverse covariance matrix is computed. Based on the assumption of sparsity (i.e., a large number of entries in the inverse covariance matrix could be set to 0 without losing much information), a set of penalty parameters are applied, forcing weak or redundant correlations among pairs of variables to zero. SICE techniques employ powerful algorithms such as the graphical lasso (Friedman *et al.*, 2008), allowing the penalty parameters to be set at different levels. The higher the penalty, the more sparse the resulting inverse covariance matrix. The process of trimming conditional dependencies (weak correlations) allows the detection of intrinsic correlations within the data.

Uses and advantages of SICE

Table 1 illustrates the uses and advantages of SICE via a comparison with several other well-known analytical approaches. First, and most importantly, SICE is an exploratory technique. In common confirmatory models, such as linear regression or Structural Equation Modeling (SEM), the model structure is specified based on domain theory, and data are used for estimating the unknown parameters and/or evaluating the model fit (Aguinis, 2004; Williams, Vandenburg, and Edwards, 2009). In contrast, in the SICE approach, data are

DOI: 10.1002/smj

Table 1. Comparison of SICE with other analytical techniques

Technique	Primary use	Research focus	Identifies relationships among variables	Identifies relationship strength	Suitable for large datasets	Suitable for datasets where $p > n$*
Linear regression	Parameter estimation and model fit	Confirmatory	Yes	Yes	Yes	No
Structural equation modelling (SEM)	Parameter estimation and model fit	Confirmatory	Yes	Yes	Yes	No
Kernel density estimation	Nonparametric probability function estimation	Exploratory or confirmatory	No	No	Yes	No
Cluster analysis	Data aggregation into groups	Mostly exploratory	No	No	Yes	No
Case study/grounded theory	Theory building and hypothesis generation	Exploratory	Yes	No	No	Yes
Qualitative comparative analysis (QCA)	Assessing multiple combinations of causal conditions	Mostly exploratory	Yes	Yes	Yes	Limited
Sparse inverse covariance estimation (SICE)	Intrinsic relationship detection and hypothesis generation	Exploratory	Yes	Yes	Yes	Yes

*p denotes the number of variables (parameters) and n denotes the number of observations.

used to specify the model structure itself, and the model structure produced is usually unique under very mild conditions (i.e., given the same penalty parameter and the same algorithm). With the recent advancement of computational and optimization techniques, this can be achieved by employing an optimization component that automatically explores the data to seek the best model, guided by a scoring mechanism that provides an index of the goodness-of-fit of each model.

This approach is therefore especially useful in the early development stage of a theory or research domain. Unexpected findings from an SICE analysis can be used to motivate the creation of new hypotheses. These hypotheses, in conjunction with both new and existing theory, can then guide the specification of confirmatory models for future research. For example, Bullmore *et al.* (2000) provide an illustration of using SICE to help revise and improve a subsequent SEM model. However, we emphasize that SICE should not be seen as an alternative to existing confirmatory analytical tools. Rather, it can be viewed as a complement and a precursor to these tools, in that it can assist in selecting appropriate model parameters and deriving more parsimonious models. In line with recent best practice recommendations (Bettis, 2012; Bettis *et al.*, 2014), researchers using SICE results to build a testable model should be sure to evaluate this model using a separate dataset or an unused portion of the existing dataset.

Second, SICE is used to detect relationships among individual pairs of variables. In contrast, alternative empirical approaches often have different primary uses and will therefore tend to be more appropriate in different situations (see Table 1). For instance, kernel density estimation is used to visualize the underlying distributions of one or more variables (Jeon and Taylor, 2012; Rawley, 2010), cluster analysis uses the collective relationships among a set of variables to aggregate units of observation into related groups (Ketchen and Shook, 1996), while qualitative comparative analysis (QCA) (Greckhamer *et al.*, 2008; Rihoux and Ragin, 2009) is used to assess multiple combinations of causal conditions on selected outcome constructs.

Third, SICE analyses can readily handle thousands of variables simultaneously, and can therefore be used with very large datasets (Friedman *et al.*, 2008). In contrast, most qualitative empirical techniques are more suited to detailed analysis of a small number of cases (Eisenhardt and Graebner, 2007). Additionally, recent developments in SICE optimization techniques also make it possible to detect intrinsic relationships in high-dimension–low-observation datasets (Huang *et al.*, 2010; Meinshausen and Bühlmann, 2006). In contrast, researchers using standard confirmatory techniques usually have to select a small portion of the available variables and establish models based on this subset (Williams *et al.*,

DOI: 10.1002/smj

2009). And, although QCA is well-suited to small-to-medium-n samples, it is limited in its ability to handle high-dimension–low-observation samples. The nature of configuration analysis in QCA is such that the number of configurations is 2^n, where n is the number of independent variables of interest (Greckhamer *et al.*, 2008). Thus, the set of logical possible configurations rises exponentially with the number of variables considered, limiting the number of variables that can be considered at once.

Limitations of SICE

SICE also has several limitations, some of which can be mitigated. First, SICE techniques are typically not able to detect whether relationships among variables are directed or causal (although see Spirtes *et al.* (2010) for possible exceptions). This reinforces our earlier point that SICE analyses should be used as a prelude to detailed theory building and subsequent hypothesis testing. Second, SICE analyses do not reveal latent relationships among variables. However, if researchers believe that there are likely to be a number of important latent constructs in a dataset, one possibility is first to perform a factor analysis to identify the latent factors and then to run SICE analyses using the latent constructs. Third, classic SICE approaches (1) assumed that all variables were continuous and conformed to normality, and (2) were designed to detect linear relationships rather than curvilinear or moderating effects. Recent extensions of SICE have also been developed to relax these assumptions in order to detect nonlinear relationships between mixed type variables (e.g., some are continuous and some are discrete (Fellinghauer *et al.*, 2013; Lee and Hastie, 2013)). In addition, an alternative approach to detecting moderating effects is to divide data into multiple categories along one or more theoretical dimensions of interest, and then run SICE analyses separately on each individual group (similar to multiple-group SEM analysis). Results can then be compared across the groups to detect any discrepancies.

In summary, the SICE approach offers strategic management researchers a powerful set of techniques for detecting intrinsic bivariate relationships within large, high-dimension–low-observation datasets. We now provide a detailed example of using SICE on the High-Performance Manufacturing (HPM) dataset.

APPLYING SICE TO A HIGH-DIMENSION–LOW-OBSERVATION DATASET

Sample

The HPM dataset was compiled by a cross-national group of researchers interested in understanding which business practices drive firm performance (see Flynn *et al.* (1997) for more detail). Plant-level cross-sectional data were collected in several waves. In this study, we used data from wave three, which comprised 12 key categories of data collected from a range of respondents. The initial sample comprised 1,705 variables and 197 plant-level observations. We then removed variables with 75 percent or more missing data, resulting in a sample of 1,691 variables. For those variables with less than 75 percent missing data, we imputed missing data using the nearest neighbor method (Wasito and Mirkin, 2005). Finally, we standardized all variables.

Stage I: exploring the intrinsic relationships among variables

Prior research has identified numerous factors that can impact firm performance, from human resource policies (e.g., Youndt *et al.*, 1996) to manufacturing practices (e.g., Flynn, Sakakibara, and Schroeder, 1995) to organizational culture (Baird, Hu, and Reeve, 2011). We therefore expected to identify relationships consistent with existing theory, but were also interested in uncovering unexpected relationships that offered the potential for generating new hypotheses.

After creating an inverse covariance matrix for the full sample, we used different levels of penalty parameters to trim conditional dependencies and discard the redundant correlations among the variables (see Figure 1). The larger the penalty value, the more sparse the graph appears, and the fewer intrinsic relationships remain (see Friedman *et al.*, 2008, for further discussion of penalty parameters). We took 1,000 equally spaced values as the input parameters and generated 1,000 corresponding sparse inverse covariance matrices. Figure 1 depicts three representative matrices within this set. The horizontal and vertical axes in each graph represent the variables from 1 to 1,691. A visible cell inside a graph represents an arc, which is a conditional dependency between the respective

DOI: 10.1002/smj

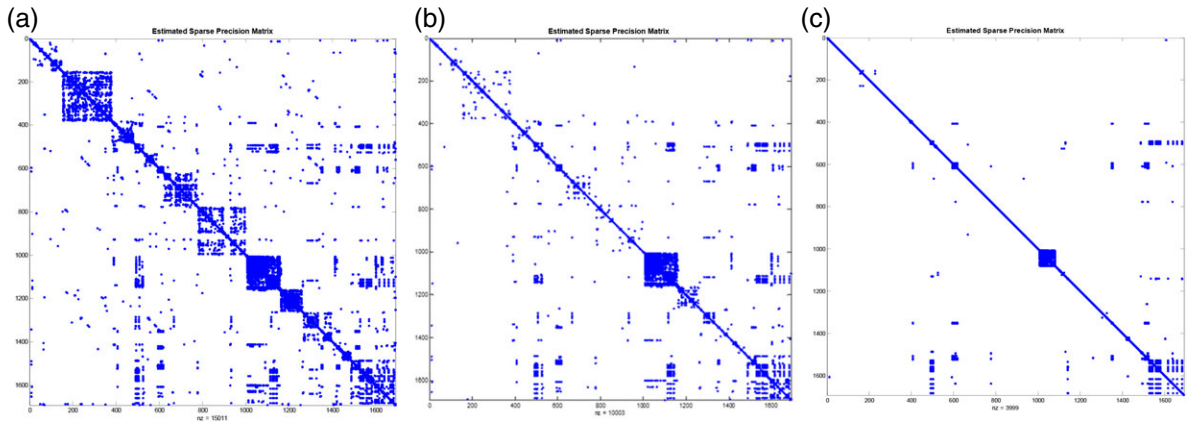

Figure 1. SICE analysis based on the full HPM dataset ($p = 1{,}691$, $n = 197$). (a) Estimated sparse precision matrix when penalty = 0.428. (b) Estimated sparse precision matrix when penalty = 0.538. (c) Estimated sparse precision matrix when penalty = 0.9087

x and y variables when conditioning on all other variables. If there is no arc, there is no intrinsic relationship between two variables. Because the matrix is symmetric, the total number of cells is equal to twice the total number of arcs in the corresponding connectivity graph. The diagonal line (top left to bottom right in each graph) represents a variable's correlation with itself and therefore has no meaning. The matrices depicted in Figure 1 contain 15,011, 10,003, and 3,999 arcs, respectively (out of 2,857,790 possible variable pairs). Table 2 shows how the variable numbers within Figure 1 map on to the corresponding variable categories from the HPM dataset. The first row of Table 2 contains the data categories, while the first column contains the data sources. For instance, variables 65–151 relate to company strategy (S) and were supplied by plant managers (PM).

Figure 1 reveals several notable findings. First, there are a number of prominent square shapes along the diagonals. This indicates that a variable within one of these regions is strongly correlated with other variables whose numbers are slightly below or above that variable. Such a finding is understandable. For our analyses, we assigned variable numbers based on the order in which a variable appeared in the HPM data, so variables belonging to the same category (e.g., IT) were positioned close to each other and thus were more likely to show high correlations.

Second, beyond the squares close to the diagonal, we also see prominent arcs in several other regions. For instance, in Figure 1(a) (penalty parameter = 0.428), there is an area of dense arcs in the region from around (500, 1,000) to around (500, 1,600) (coordinates based on X, Y axes). Based on Table 2, this suggests that variables related to the information technology categories are intrinsically related to a firm's manufacturing strategy. An alternative way to explore Figure 1(a) would be to look for prominent arcs along the path of (17–29), (379–382), and (1,489–1,526). These ranges represent firm performance from the perspective of plant managers, quality managers, and accounting managers, respectively. Any arcs along these three paths would represent intrinsic relationships with firm performance. Of course, these examples only provide a high-level view of relationship patterns across broad activity domains. Researchers could then drill down to the variable level in order to guide the development of specific hypotheses.

One possible next step in multi-respondent data such as ours is to focus the analysis on data from a single source. Because strong correlations among multiple sources responding to the same items could potentially complicate our results, we selected only those data provided by supervisors (SP). We ran a new SICE analysis based on this subsample ($p = 238$, $n = 197$). Figure 2 represents the patterns of estimated sparse precision matrices under three different penalty values, while Table 3 shows the corresponding variable categories. We then used the matrix from Figure 2(c) to generate a network diagram (see Figure 3). This network includes those nodes in the subsample associated with one or more arcs. Figure 3 shows that several nodes were highly central in the network. For instance, node 75 reflects supervisors' responses to the item "We work

DOI: 10.1002/smj

Table 2. Mapping of spaces on matrices in Figure 1 and corresponding variable categories[a]

Data Source	C	E	G	H	I	J	Q	S	T	M	N	P
AC			1,489–1,526[b]									
DL				1,162–1,187			1,188–1,254	1,255–1,260				
HR		1,404–1,409		1,410–1,486	1,487–1,488							
IS	1,006–1,161											
PC		524–527				528–589						
IM		1,261–1,266		1,267–1,272		1,273–1,333	1,380–1,391	1,392–1,394	1,395–1,403			1,334–1,379
PD									1,527–1,554		1,555–1,691	
PE		590–613					638–658	659–736	737–774	614–637		
PM		1–16	17–29	30–52			53–64	65–151	152–152			
QM			379–382	383–388			389–486	487–523				
SP				153–218		219–279	327–354	355–369	370–378	280–300		301–326
PS		775–779		780–841			889–894	895–989	990–1,005	842–862		863–888

[a] Data categories: information systems/information technology (C), business environment (E), goals/performance indicators (G), human resources (H), improvements (I), just-in-time practices and theory of constraints (J), quality (Q), strategy (S), technology/mass customization (T), maintenance (M), new product development (N) supply chain (P).
[b] Information source: plant accounting manager (AC), direct labor (DL), human resources manager (HR), information systems manager (IS), production control manager (PC), inventory manager (IM), member of product development team (PD), process engineer (PE), plant manager (PM), quality manager (QM), supervisor (SP) plant superintendent (PS). Multiple sources provided data in each category.

as a partner with our customers" (HR category). This node has intrinsic relationships with a range of nodes from different domains, including customer involvement (node 191), supply chain planning (node 179), and, interestingly, supplier lead time (node 167). A possible hypothesis might therefore be "Supplier lead time is negatively associated with firm–customer relationship strength." It would be then up to the researcher to build a compelling theoretical explanation for this hypothesis (perhaps by drawing in part from the "learning by supplying" literature (Alcacer and Oxley, 2014)), and subsequently testing the hypothesis in a new dataset.

Stage II: comparison of intrinsic relationships across high- and low-performing groups

We then examined whether there were differences in the relationship patterns among variables across high- and low-performing firms. We first divided our data into two groups based on whether a firm was on an upward or downward trajectory in annual sales. The high-performing group comprised 44 firms, while the low-performing group comprised 51 firms (the remaining firms had missing data or reported no change in annual sales). Figure 4 depicts the resulting set of estimated sparse inverse covariance matrices for these two groups. Figure 4(a) shows the relationship patterns when we created matrices with 10,000 arcs (using a penalty parameter of 0.723 for the high-performing group and 0.729 for the low-performing group). Figure 4(b) shows the relationship patterns when we created matrices with 5,000 arcs (using a penalty parameter of 0.949 for the high-performing group and 0.945 for the low-performing group).

In these analyses, we kept the number of arcs constant to ensure that the sparsity (i.e., the number of relationships that were set to zero) of comparable matrices was identical, which allowed us to contrast the structural differences in the inverse covariance matrices of the two datasets. This is preferable to the alternative approach of holding the penalty parameter constant and letting the number of arcs vary, because such an alternative would preclude meaningful comparisons across matrices. For different datasets (such as high- vs. low-performing groups), different inverse covariance matrices will exist. Therefore, applying the same penalty parameter to different datasets will cause inconsistent effects. For example, a particular parameter (say, 0.1) may be insignificant to increase the sparsity for

DOI: 10.1002/smj

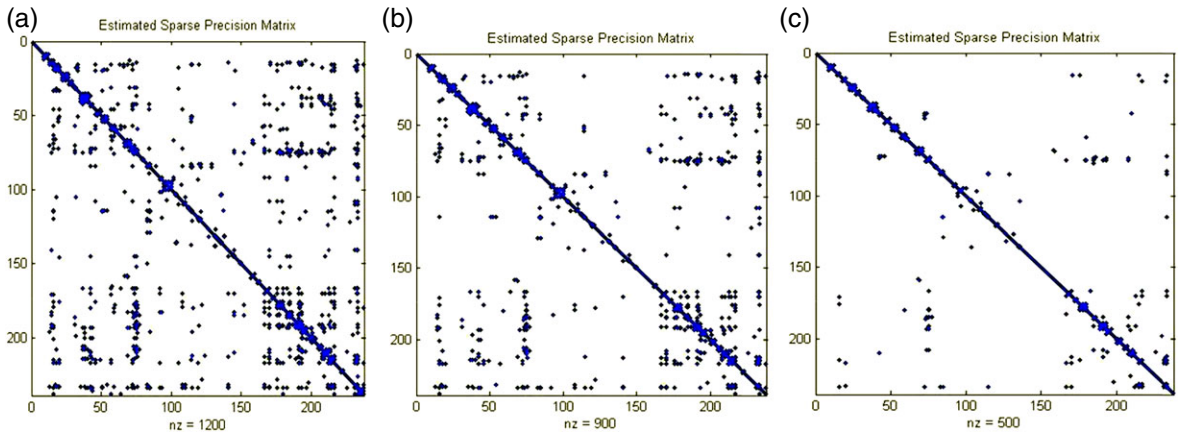

Figure 2. SICE analysis based on supervisor responses only ($p = 238$, $n = 197$). (a) Estimated sparse precision matrix when penalty = 0.521. (b) Estimated sparse precision matrix when penalty = 0.544. (c) Estimated sparse precision matrix when penalty = 0.591.

Table 3. Mapping of spaces on matrices in Figure 2 and corresponding variable categories

Categories[a]	H	J	Q	S	T	M	P
Variable range	13–78	79–139	187–214	215–229	230–238	140–160	161–186

[a] Data categories: human resources (H), just-in-time practices and theory of constraints (J), quality (Q), strategy (S), technology/mass customization (T), maintenance (M), supply chain (P).

one dataset but have a significant effect in another dataset. Hence, we do not recommend this alternative approach.

Analysis of these graphs revealed little variation in relationship density patterns as a function of firm performance. Many variables, such as respondents' views on the commercial success of new products, were present in both subgroups. We therefore conducted a more fine-grained analysis by excluding all variables with common arcs in both high-performing and low-performing groups and rerunning our SICE analyses using the remaining variables (N = 1,256). Excluding common arcs allowed us to eliminate some of the shared statistical noise in the datasets and thus make the relationships among the remaining variable pairs more salient and detectable. This additional analysis revealed that there were indeed some notable cross-group differences. For instance, arcs linked with several variables—such as employees' attitudes toward firm-level outcomes (e.g., "most of our employees try to help our organization achieve its goals") and whether a firm conducted technical analyses of major breakdowns—were prominent in the high-performing firms but not the low-performing ones. Researchers could use findings such as these to help develop contingent hypotheses predicting organizational performance.

One possible concern when comparing differences in the correlation pattern across groups, especially in light of the very large number of potential correlations involved, is that such differences might simply be due to random chance. Although there is currently no specific, widely-used test to address this issue in the SICE literature, researchers can increase the likelihood of particular relationships being substantive by (1) using higher penalty parameters, ensuring that selected relationships reflect moderate or large effect sizes (Kelley and Preacher, 2012), and (2) formally testing the identified relationships in separate datasets.

Stage III: exploration of intrinsic relationships among latent factors

Finally, we conducted a further SICE analysis after aggregating the HPM data into a number of latent factors. We began by performing a principle components analysis on the existing 1,691 variables, which resulted in a reduced sample of 286 components (i.e., $p = 286$; $n = 197$). Similar to the corresponding procedure at the individual variable level, this

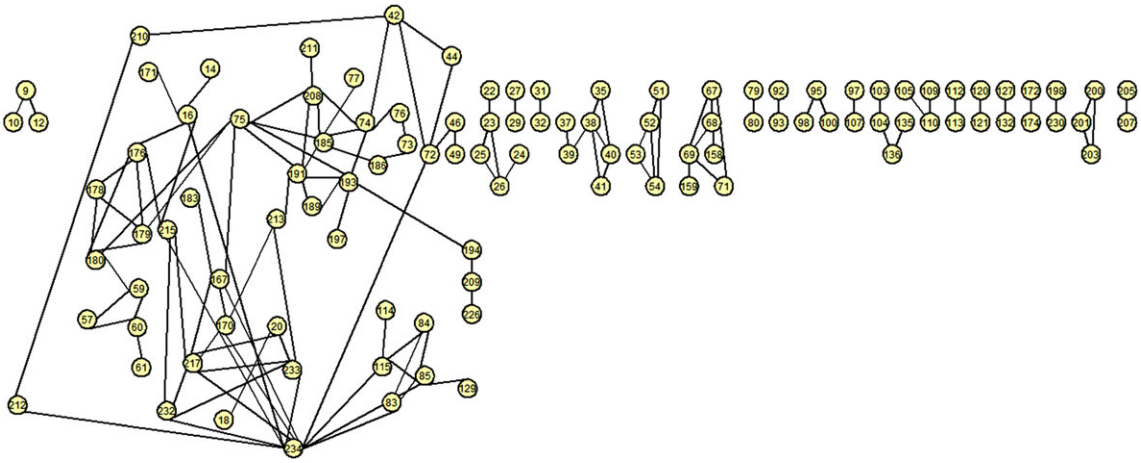

Figure 3. Network structure based on the estimated sparse precision matrix depicted in Figure 2(c) ($p = 238$, $n = 197$)

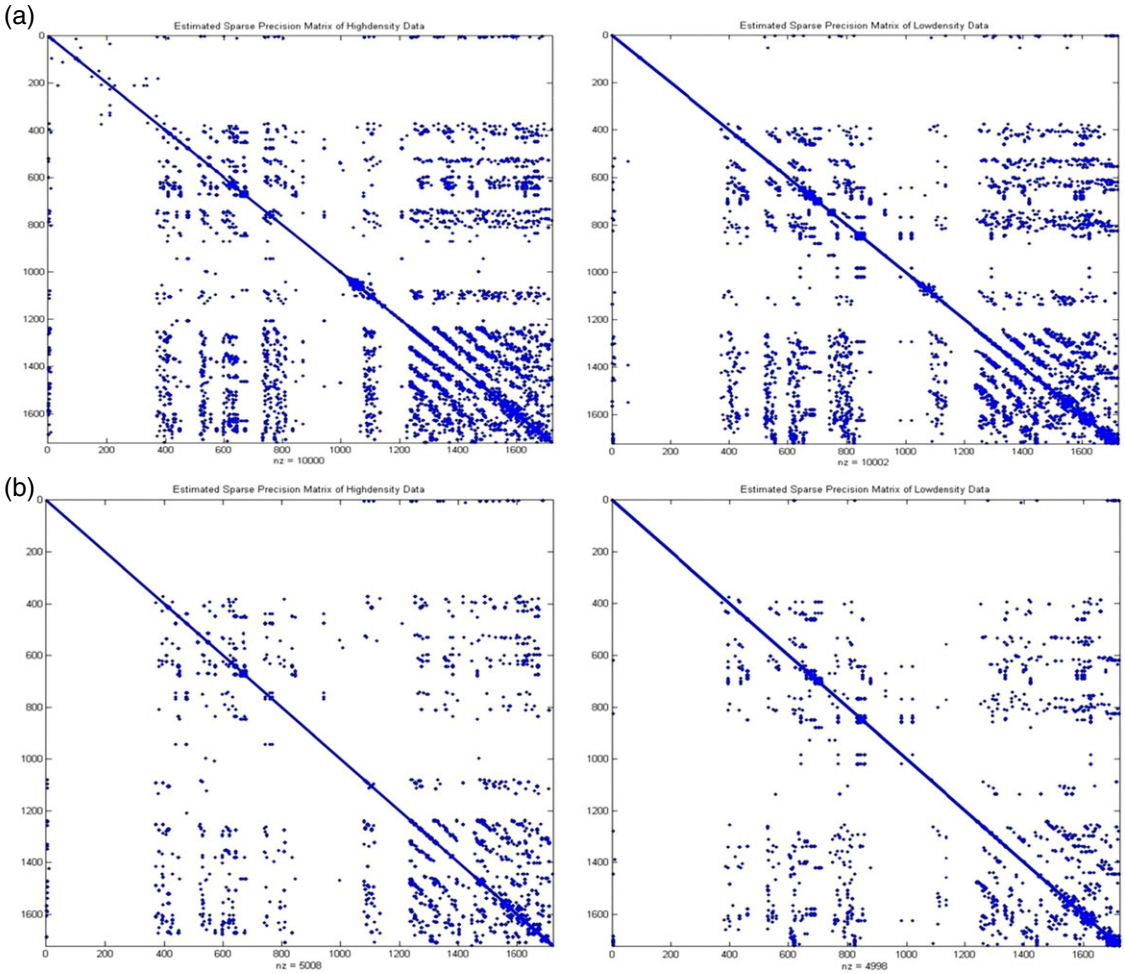

Figure 4. Comparison of estimated sparse precision matrices across high-performing and low-performing groups of firms. (a) Number of arcs per graph = 10,000 (high-performers: penalty = 0.723; low-performers: penalty = 0.729). (b) Number of arcs per graph = 5,000 (high-performers: penalty = 0.949; Low-performers: penalty = 0.945).

DOI: 10.1002/smj

approach generated several notable findings. For example, there was a prominent arc between a component denoting certain internal HR practices of a firm and a component denoting how well a firm's employees worked with their external suppliers. When we drilled down, we found that the variable exerting most influence in the former component was a firm's employee screening practices. Human resource practices, such as employee screening, are a reflection of a firm's internal policies, while an employee's interaction with a supplier is more likely to be guided by a firm's vendor management policy (i.e., an external policy). These two types of policies can differ significantly. For example, Hewlett-Packard (HP) places a heavy emphasis on internal employee collaboration. Teamwork is a central component of the "HP Way" (Kotelnikov, 2009) and is a part of employees' annual performance evaluation. In contrast with HP's more collaborative internal policies, though, the firm also has a reputation for keeping its external vendors at arm's length (Bussey, 2011). Thus, within HP at least, there appears to be some separation between intra-firm and interfirm collaborative practices. However, our SICE results suggest that these two types of practices may be linked in firms more generally and that intra-firm collaborative orientations may be linked with interfirm collaborative dynamics.

To explore the implications of this result further, we tested a similar relationship using a different set of actors. Our idea was that, if this association holds for relationship dynamics between a buyer firm and a supplier firm, it should also hold for relationship dynamics between a firm and its end customers. We tested this idea using another segment of the HPM dataset and a partial least squares (PLS) analysis. The results of this analysis provided support for our general idea that intra-firm and interfirm collaborative practices are linked (full analyses available on request). Although this is a simple example (and we recommend validating such a finding in an entirely separate dataset), it illustrates how researchers can use SICE to generate new hypotheses in one context and then test those and other related extensions in a different context.

DISCUSSION

In this paper, we introduced to the strategic management field a set of exploratory quantitative techniques known as Sparse Inverse Covariance Estimation (SICE). These techniques, which have been previously used mostly in the biological and physical sciences (e.g., Huang *et al.*, 2010; Valdes-Sosa *et al.*, 2005), offer strategy researchers a rigorous means for identifying the fundamental relationships among a network of variables, thereby generating novel and interesting hypotheses. We discussed the uses, strengths, and limitations of SICE and provided a detailed example of its application techniques to the High Performance Manufacturing (HPM) dataset. As noted above, we also direct interested readers to the Online Supplemental Material associated with this paper for a detailed discussion of the conceptual foundations of SICE, information on model structure identification and evaluation, a mathematical proof, and instructions for implementation of SICE, including MATLAB programming codes.

Our study has implications for several avenues of existing work both within and outside strategic management. First, we believe that SICE techniques can help to build a bridge between strategic management and "big data" (George, Hass, and Pentland, 2014). The notion of using quantitative analyses to identify conceptual relationships and generate hypotheses is timelier now than ever before. Researchers interested in studying firms and their business practices—the preferred units of analysis within strategic management—are experiencing an exponential increase in data availability (George *et al.*, 2014). In addition to the thousands of firm-level variables available via established databases, such as COMPUSTAT, CRSP, KLD, and Worldscope, large and growing pools of data provide millions more potential measures of interest, including an enormous range of company-, customer-, and stakeholder-generated measures of firms' actions, processes, and outputs (McKinsey Global Institute, 2011).

Current discussions of the use of big data in organizations address a multiplicity of issues, from its promise for providing answers to narrow customer-level purchasing decisions all the way up to its potential for transforming firm-wide operations and strategy (LaValle *et al.*, 2011). However, the opportunities of big data lie less in its quantity and more in the quality of its analysis, with a recent survey indicating that more than 60 percent of business executives felt that their organizations already had more data than they could use (LaValle *et al.*, 2011). Researchers dealing with big data

DOI: 10.1002/smj

face similar challenges, especially those relying on established metrics such as p-values to determine whether a conceptual relationship exists (George *et al.*, 2014). In very large datasets, the question is not so much *whether* a particular relationship exists (a sufficiently high number of observations will make almost all relationships significant at conventional values) as *which* are the most important relationships (Mitchell and Leiponen, 2015). SICE techniques are especially useful for answering this question, thereby providing a crucial step toward developing theory to explain why particular relationships exist in the first place.

Second, and relatedly, although we illustrated the use of SICE with the HPM dataset only, we want to emphasize that these techniques are likely to be of interest to researchers working in a wide range of subdomains within strategic management. For instance, strategic leadership researchers (Finkelstein, Hambrick, and Cannella, 2009) now have access to a growing amount of information on senior executives' backgrounds, experiences, values, and preferences because of the widespread and increasing use of unobtrusive measures to measure executives' individual difference characteristics (e.g., Petrenko *et al.*, 2015). Combining this with, say, governance data from BoardEx, compensation data from Execucomp, and strategic behavior and firm outcome data from COMPUSTAT, researchers could easily find themselves dealing with thousands of potential variables. The increasing availability of granular data on a wide variety of firm-level architectures and routines (e.g., Joseph and Ocasio, 2012) suggests that strategy process researchers must often deal with a related issue at the level of business practices, while researchers dealing with mergers and acquisitions, innovation, and corporate social responsibility (e.g., Godfrey, Merrill, and Hansen, 2009; Klingebiel and Rammer, 2014; Sears and Hoetker, 2014) all face similar challenges. In such situations, standard confirmatory research focuses on the impact of one or more variables within a small subset of the data. In contrast, exploratory SICE techniques shift the focus to the intrinsic relationships within the dataset as a whole, with the goal of identifying novel associations and hypotheses. Finally, many strategy scholars are especially interested in the most powerful corporate actors. Thus, scholars often restrict their focus to the S&P 500 or 1,500, but have access to thousands of potentially informative predictors. SICE techniques are well suited to identifying the important underlying relationships among variables in exactly these types of high-dimension–low-observation contexts.

REFERENCES

Aguinis H. 2004. *Regression Analysis for Categorical Moderators*. Guilford: New York.

Ahmed A, Xing EP. 2009. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences* **106**: 11878–11883.

Alcacer J, Oxley J. 2014. Learning by supplying. *Strategic Management Journal* **35**: 204–223.

Baird K, Hu KJ, Reeve R. 2011. The relationships between organizational culture, total quality management practices and operational performance. *International Journal of Operations & Production Management* **31**: 789–814.

Bascle G. 2008. Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization* **6**: 285–327.

Bettis RA. 2012. The search for asterisks: compromised statistical tests and flawed theories. *Strategic Management Journal* **33**: 108–113.

Bettis RA, Gambardella A, Helfat C, Mitchell W. 2014. Quantitative empirical analysis in strategic management. *Strategic Management Journal* **35**: 949–953.

Bullmore E, Horwitz B, Honey G, Brammer M, Williams S, Sharma T. 2000. How good is good enough in path analysis of fMRI data? *NeuroImage* **11**: 289–301.

Bussey J. 2011. Measuring the human cost of an iPad made in China. Wall Street Journal, June 3. http://www.wsj.com/articles/SB10001424052702304563104576361232998099752?alg=y

Dempster AP. 1972. Covariance selection. *Biometrics* **28**: 157–175.

Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. 2004. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**: 196–212.

Eisenhardt KM, Graebner ME. 2007. Theory building from cases: opportunities and challenges. *Academy of Management Journal* **50**: 25–32.

Fan J, Lv J, Qi L. 2011. Sparse high-dimensional models in economics. *Annual Review of Economics* **3**: 291–317.

Fellinghauer B, Bühlmann P, Ryffel M, von Rhein M, Reinhardt JD. 2013. Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis* **64**: 132–152.

Finkelstein S, Hambrick DC, Cannella AA. 2009. *Strategic Leadership: Theory and Research on Executives, Top Management Teams, and Boards*. Oxford University Press: New York.

Flynn BB, Sakakibara S, Schroeder RG. 1995. Relationship between JIT and TQM: practices and performance. *Academy of Management Journal* **38**: 1325–1360.

Flynn BB, Schroeder RG, Flynn EJ, Sakakibara S, Bates KA. 1997. World-class manufacturing project:

overview and selected results. *International Journal of Operations & Production Management* **17**: 671–685.
Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**: 432–441.
Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**: 1–22.
George G, Hass MR, Pentland A. 2014. Big data and management. *Academy of Management Journal* **57**: 321–326.
Gioia DA, Chittipeddi K. 1991. Sensemaking and sensegiving in strategic change initiation. *Strategic Management Journal* **12**: 433–448.
Godfrey PC, Merrill CB, Hansen JM. 2009. The relationships between corporate social responsibility and shareholder value: an empirical test of the risk management hypothesis. *Strategic Management Journal* **30**: 425–445.
Greckhamer T, Misangyi VF, Elms H, Lacey R. 2008. Using qualitative comparative analysis in strategic management research. *Organizational Research Methods* **11**: 695–726.
Huang S, Li J, Sun L, Ye J, Fleisher A, Wu T, Chen K, Reiman E, Alzheimer's Disease NeuroImaging Initiative. 2010. Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *NeuroImage* **50**: 935–949.
Jeon J, Taylor JW. 2012. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association* **107**: 66–79.
Jones DT, Buchan DWA, Cozzetto D, Pontil M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**: 184–190.
Joseph J, Ocasio W. 2012. Architecture, attention, and adaptation in the multibusiness firm: General Electric from 1951 to 2001. *Strategic Management Journal* **33**: 633–660.
Kelley K, Preacher KJ. 2012. On effect size. *Psychological Methods* **17**: 137–152.
Ketchen DJ, Shook CL. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal* **17**: 441–458.
Kilduff M. 2006. Publishing theory. *Academy of Management Review* **31**: 252–255.
Kilduff M, Mehra A, Dunn MB. 2011. From blue sky research to problem solving: a philosophy of science theory of new knowledge production. *Academy of Management Review* **36**: 297–317.
Kim N, Tikves S, Wang Z, Githens-Mazer J, Davulcu H. 2013. MultiScale modeling of Islamic organizations in UK. In *International Conference on Social Computing*, Stanford, CA, USA, 13–18.
Klingebiel R, Rammer C. 2014. Resource allocation strategy for innovation portfolio management. *Strategic Management Journal* **35**: 246–268.
Kotelnikov V. 2009. The Hewlett-Packard way: sustaining competitive advantage by managing critical opposites. 1000 Ventures.
Kuhn TS. 1996. *The Structure of Scientific Revolutions* (3rd edn). University of Chicago Press: Chicago, IL.
Lakatos I. 1970. Falsification and the methodology of scientific research programs. In *Criticism and the Growth of Knowledge*, Lakatos I, Musgrave A (eds). Cambridge University Press: Cambridge, UK; 91–196.
LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N. 2011. Big data, analytics, and the path from insights to value. *MIT Sloan Management Review*, Winter Special Issue **52**: 21–31.
Lee JD, Hastie TJ. 2013. Structure learning of mixed graphical models. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 388–396.
Mahoney JT, McGahan AM. 2007. The field of strategic management within the evolving science of strategic organization. *Strategic Organization* **5**: 79–99.
McKinsey Global Institute. 2011. *Big Data: The New Frontier for Innovation, Competition, and Productivity*. McKinsey & Company: Lexington, KY.
Meinshausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**: 1436–1462.
Mitchell W, Leiponen A. 2015. Virtual special issue on innovation, intellectual property, and strategic management. *Strategic Management Journal* , DOI: 10.1002/smj.2282, in press.
Myers SA, Leskovec J. 2010. On the convexity of latent social network inference. In *Neural Information Processing Systems (NIPS) Annual Conference Proceedings*, Vancouver, Canada, 1741–1749.
Petrenko OV, Aime F, Ridge J, Hill A. 2015. Corporate social responsibility or CEO narcissism? CSR motivations and organizational performance. *Strategic Management Journal* , DOI: 10.1002/smj.2348, in press.
Popper KR. 1959. *The Logic of Scientific Discovery*. Routledge: London, UK.
Rawley E. 2010. Diversification, coordination costs, and organizational rigidity: evidence from microdata. *Strategic Management Journal* **31**: 873–891.
Rihoux B, Ragin CC (eds). 2009. *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. Sage: Thousand Oaks, CA.
Sears J, Hoetker G. 2014. Technological overlap, technological capabilities, and resource recombination in technological acquisitions. *Strategic Management Journal* **35**: 48–67.
Semadeni M, Withers MC, Certo ST. 2014. The perils of endogeneity and instrumental variables in strategy research: understanding through simulations. *Strategic Management Journal* **35**: 1070–1079.
Spirtes P, Glymour C, Scheines R, Tillman R. 2010. Automated search for causal relations: theory and practice. In *Heuristics, Probability, and Causality: A Tribute to Judea Pearl*, Dechter R, Geffner H, Halpern JY (eds). College Publications: London, UK; 467–506.
Toh H, Horimoto K. 2002. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* **18**: 287–197.
Valdes-Sosa PA, Sanchez-Bornot JM, Lage-Castellanos A, Vega-Hernandez M, Bosch-Bayard J, Melie-Garcia

L, Canales-Rodriguez E. 2005. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B* **360**: 969–981.

Wasito I, Mirkin B. 2005. Nearest neighbor approach in the least-squares data imputation algorithms. *Information Sciences* **169**: 1–25.

Weiss Y, Freeman WT. 2001. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation* **13**: 2173–2200.

Williams LJ, Vandenburg RJ, Edwards JR. 2009. Structural equation modeling in management research: a guide for improved analysis. *Academy of Management Annals* **3**: 543–604.

Youndt MA, Snell SA, Dean JW, Lepak DP. 1996. Human resource management, manufacturing strategy, and firm performance. *Academy of Management Journal* **39**: 836–866.

Zhang W, Fung P. 2013. Sparse inverse covariance matrices for low resource speech recognition. *IEEE Transactions on Audio, Speech and Language Processing* **21**: 659–668.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix S1. The use of sparse inverse covariance estimation for relationship detection and hypothesis generation in strategic management.