SeDFAM: Semiconductor Demand Forecast Accuracy Model *

Metin Çakanyıldırım and Robin O. Roundy

www.orie.cornell.university/~metin/semicond

School of Operations Research and Industrial Engineering
Cornell University

November 24, 1999

Abstract

In the semiconductor industry, many critical decisions are based on demand forecasts. However, these forecasts are subject to random error. In this paper, we lay out a scheme estimating the variance and correlation of forecast errors (without altering given forecasts) and modeling the evolution of forecasts over time. Our scheme allows correlations across time, products and technologies. It also addresses the case of nonstationary errors due to ramps (technology migrations). It can be used to simulate chip demands for production planning / capacity expansion studies.

*Technical Report 1230, SORIE. Cornell University 14853, NY.

1 Introduction

In this paper, we will attempt to quantify forecast errors, i.e., the differences between forecasts and actual demands for semiconductors. Some of the most important, difficult and risky decisions made by semiconductor companies are based on demand forecasts. Capacity acquisition and deployment are prime examples. We take a given set of demand forecasts, model the forecast error as a vector-valued random variable, and estimate its distribution. The estimate is derived from historical forecasts and demand data. Our primary goal is to quantify the risks associated with capacity acquisition decisions, and with a variety of other forecast-based decisions.

Generation of forecasts is outside the scope of this study. Semiconductor demand forecasting is a complex process. In this process, forecasters use many sources of information. The process is difficult to quantify and represent mathematically. Therefore, we do not propose an alternative forecasting scheme.

The study of forecast quality is particularly relevant for the semiconductor industry because high volatility makes forecasting challenging. Moreover, it is believed in the industry that forecast accuracy is getting worse, both because product life cycles are shortening and because line-widths are shrinking (Çakanyıldırım and Roundy [1]). Line-width shrinkage often makes it harder to sell chips produced under an old technology with wider widths unless substantial discounts are made. Thus, demand for a certain line-width of a certain product often resembles that of a style good which will not be demanded after a certain point in time.

Within the context of forecasting, semiconductor product families can be divided roughly into two groups. Families in the first group have persistent demand over extended periods of time, such as memory, ASICs, CPUs, controller chips for desktop printers, etc. As time advances these families evolve, and they are migrated to finer and finer line widths, but the overall demand for the product family continues. For these product families, the requirements for manufacturing capacity is a function of overall family demand, of line width migrations, and of a variety of other technological factors, all of which need to be forecasted.

Product families in the second group are just coming into existence, and consequently have no historical data that can be used in forecasting. For these families methodologies that were developed for style goods and by Meixell and Wu [2] are applicable. This paper focuses on semiconductor families that have persistent demand over extended periods of time.

A primary concern will be the evolution of forecasts as time passes. Consider forecasts of domestic PC sales in December 2000. Suppose we start forecasting at the beginning of January 2000 and update the forecast every month. At the beginning of December 2000, we will have produced 12 forecasts for December sales. Actual sales figure for December will be available at the end of December. This figure can be thought as a forecast with zero error. Putting these 13 numbers, all being forecasts for December sales, into chronological order, a time series is obtained. This series shows how forecasts for December sales evolve over time from a highly uncertain forecast in January to a sharply accurate one in December. The reason for forecast evolution is the additional information becoming available month by month. Such information can be about sales, customer behaviors, the status of new manufacturing technologies and promotion campaigns. The resolution rate of forecast uncertainty is a function of the amount of additional information acquired as time passes.

2 Literature Survey

There is a huge time-series literature on methods to generate demand forecasts (see Hamilton [3]). Since we investigate forecast errors for a style good (semiconductors) as opposed to forecast generation, we mention a few, relatively recent, papers that study forecasting, especially for new products or style goods. Mahajan and Wind [4] survey the new product forecasting models. Murray and Silver [5] represent the demand of a style good as a binomial random variable where the number of potential buyers is constant but the probability of a customer purchase is updated using past sales. Chang and Fyffe [6] visualize monthly demands as a fixed fraction of the total demand for a style good. Total demands are modified as monthly sales are revealed.

In forecasting new product demands, the Bass function [7] is often used. Norton and Bass [8] model diffusion of a new product (demand migration) in the markets. Kurawarwala and Matsuo [9] study seasonal PC demands.

In both models parameters of the Bass function are updated periodically.

Another stream studies forecast revisions with partially observed (demand) data. The main driving idea in Guerrero and Elizondo [10], Kekre et al. [11], and Bodily and Freeland [12] is presuming that the forecasted quantity is revealed in steps (over time). Moreover, partial observations can be used to forecast the whole quantity. For example, in one of the models it is assumed that the ratio of orders received up to a certain time to the whole demand is approximately constant. After the proportionality constant is estimated, forecasts are readily generated from the partially observed demands.

A relatively early example of updating forecasts in a Bayesian manner is Azoury [13]. It assumes that demand has a particular density and updates parameters of that density by conditioning on previous demands. The idea of updating forecasts in a Bayesian fashion through leading indicator products is introduced by Meixell and Wu in [2]. In this scheme, forecaster observes the demand for a particular product and revises forecasts for other products whose demands are strongly correlated with those of the particular product. For new products, and for products that are being phased out, the general approaches that have been developed within this scheme, and those for style good forecasting, are applicable. The scheme developed by Meixell and Wu outputs demand scenarios instead of forecasts, thus it constitutes an example of the *contemporary practice of forecasting with scenarios* argued by Bunn and Salo [14]. Angelus et al. [15] ties the current period's demand to the last period's demand through a random multiplier whose expected value is greater than one. Works of Meixell and Wu, and Angelus are done on semiconductor demands and so are particularly relevant. However, they primarily aim to generate demand scenarios/forecasts whereas our primary goal is to evaluate the quality of the forecasts.

As far as we know, the first work that attempted to understand forecast variances is Graves, Meal, Dasu and Qui [16] where the forecasting process is treated as a black box. However, [16] assumes that forecasts are serially independent. Heath and Jackson [17], unaware of [16], propose a more general model allowing serial dependence and estimate the covariance matrix of stationary demand forecasts. That matrix is also used to simulate demand forecasts. In recent studies, Güllü [18] and Toktay [19] have investigated the value of incorporating this technique into production planning and inventory holding, respectively. Graves, Kletter and Hetzel [20] also use Heath-Jackson framework to study production smoothing and safety stock holding tradeoffs. They show how to linearly convert forecast updates into production schedule updates so that a measure of production smoothing is minimized subject to an upper bound on a measure of safety stock.

We focus on semiconductor product families that have persistent demand over extended periods of time, but which are affected by successive waves of technological innovation and improvement. The level of detail in our models is driven by a desire to support the acquisition of manufacturing capacity, and related decisions. At this level of detail demands become nonstationary with technology improvements. We propose a forecast evolution model that handles nonstationarity and that is a result of ideas blended from Heath-Jackson framework and style goods forecasting. Our primary goal is the estimation of variances and covariances of forecast errors affecting different products in different time periods. We have also created a tool that can be used to simulate the manner in which forecasts evolve over time.

3 A Forecast Evolution Model

We group part numbers into product families at the highest level and into products at a finer level. Chips belonging to the same functional category are put into the same product family. For example *memory chips* constitute a product family. Chips of the same product family are further grouped according to technology (e.g. CMOS 12). Let p and tec denote a generic product family and a technology respectively. The product family p and the technology tec define a product (p, tec) uniquely for our purposes. An example of a product defined as such is (Memory, CMOS 12).

In the semiconductor industry, within each product family, demand for a given technology dies out and is replaced by demand for a newer technology. We call this process *migration of product families*. The S-curve (p, tec) represents forecasted demand for the family p and for the technology tec plus demands for all technologies newer than tec . In Figure 1, a piece of the S-curve (Memory, CMOS 8), and S-curves (Memory, CMOS 10) and (Memory, CMOS 12) are

shown . Due to migration of product families, one often expects those curves to be nondecreasing. The vertical distances between consecutive S-curves is the demand for memory chips with a single technology.

Figure 1: Technology migration in a product family

In addition to working with absolute quantities on the vertical axis of Figure 1, the ratio of those absolute figures to the total product family demand will be of interest. Let $d_{s,t}^{p,tec}$ be the demand of product family p and technology tec , forecasted from period s for period t . From now on, we use the phrase *from s for t* to refer to forecasts made in period s for demands to be realized in period t . When $s = t$, the quantity in question is no longer a forecast but an observation. Let $cd_{s,t}^{p,tec}$ be the (cumulative) demand for all technologies newer than or equal to tec of family p , from s for t . Furthermore, define $d_{s,t}^p$, the *forecast for family p* , and $f_{s,t}^{p,tec}$, the *fractional forecast*, as

$$\begin{aligned}
 d_{s,t}^p &= \sum_{tec} d_{s,t}^{p,tec} \\
 cd_{s,t}^{p,tec_o} &= \sum_{tec \text{ is } tec_o \text{ or newer}} d_{s,t}^{p,tec} \\
 f_{s,t}^{p,tec} &= \frac{cd_{s,t}^{p,tec}}{d_{s,t}^p}.
 \end{aligned} \tag{1}$$

Note that $f_{s,t}^{p,tec}$ can be easily calculated from the demand forecast data.

A word of warning is in order for the units of family demands $d_{s,t}^p$. In the industry, it is customary to use number of wafers as units. However, when technology shrinkages change the number of chips per wafer, demand measured in terms of wafers is misleading. If a new memory chip stores more data and there is a constant demand for data storage, demand for chips will go down. When aggregating data, units should be chosen to minimize the impact of technology changes on total family demand.

For simplicity we discuss demands with no trend or seasonality. For the case where trend or seasonality do not affect the forecast error variances, Appendix A illustrates how to apply our approach. If trends or seasonality affect the variances, demands may need to be detrended (see [17]).

We make two critical independence assumptions and discuss them in the context of semiconductor demands:

(I1): *Product family demands and fractional demands are independent.* Family demands are primarily driven by the market (exterior forces). Contrary to that, fractional demands are driven mainly by the company's technology (interior forces). In agreement with this interior-exterior forces distinction, Meixell and Wu [2] (see section 2.2) also make an assumption similar to (I1).

(I2): *Random variables which we observe as $f_{s,t}^{p,tec_1}$ and $f_{s,t}^{p,tec_2}$ are independent for $tec_1 \neq tec_2$.* In other words shifts in different S-curves of Figure 1 are independent. An example will clarify and motivate the assumption. Let A, B denote the fractional demands for technologies 8 and 10 of the same product family at a particular time, respectively. Then the fractional cumulative demands are A+B, and B for 8 and 10, respectively. If the forecast

for B increases (corresponding to faster migration than estimated), then the forecast for A will probably decrease, because demand for B will replace demand for A. However, it is not clear how A+B will behave. No indication of strong correlations between different ramps has arisen in our interviews with people in the semiconductor industry (see [1]). Thus, assuming independence of A+B and B appears to be justifiable.

Let L be the average length of a technology ramp, i.e. the average length of the S-curves in Figure 1. We use the phrase *age of a ramp* to refer to the duration between the start of the ramp and the current time. Define a nondecreasing ramp function $R : [0, L] \rightarrow [0, 1]$, which maps the age of a typical ramp to the percentage of product family demand achieved. $\hat{R}$ is obtained by fitting a curve to all the S-curves available while requiring that $\hat{R}(0) = 0$, that $\hat{R}(L) = 1$. Let $\delta_{s,t}^{p,tec}$ be the *perceived age forecast* from s for t , where

$$\delta_{s,t}^{p,tec} = \hat{R}^{-1}(f_{s,t}^{p,tec}) \quad (2)$$

$\delta_{s,t}^{p,tec}$ is deduced from the fractional forecasts so it is different from the actual age of the ramp. Also $\delta_{s,t+1}^{p,tec}$ is not necessarily equal to $\delta_{s,t}^{p,tec} + 1$. If the ramp up during period t is forecasted to be faster than usual, then $\delta_{s,t+1}^{p,tec} > \delta_{s,t}^{p,tec} + 1$.

4 A Stochastic Formulation

We focus on a single product family and technology so that the p and tec indices on $d_{s,t}^p$, $f_{s,t}^{p,tec}$ and $\delta_{s,t}^{p,tec}$ can be dropped in this section. Note that the family demand forecast $d_{s,t}$, the fractional forecast $f_{s,t}$ and the perceived ramp age forecast $\delta_{s,t}$ are all observed at time s and they are deterministic in periods r , $r \geq s$. In periods r , $r < s$, they could be viewed as random variables because more information will be obtained between period r and period s when they are produced by the forecaster. Forecast evolution studies this incremental transition of forecasts from random variables to deterministic numbers. From now on, we will use capital letters for random variables and small letters for observations from those random variables.

In the remainder of this section, we will discuss only the evolution of perceived ramp age forecasts $\delta_{s,t}$. The evolution of product family demand forecasts is treated exactly the same way: it suffices to replace $\delta_{s,t}^{p,tec}$ with $d_{s,t}^p$ and $u_{s,t}^{p,tec}$ with $v_{s,t}^p$ in the current and the next section. All the assumptions we make in these two sections apply to product family demands and perceived ramp ages. The perceived age forecast evolution is based on updates in forecasts, which are calculated as

$$u_{s,t} = \delta_{s,t} - \delta_{s-1,t}. \quad (3)$$

Let $\mathfrak{S}_s$ be the information available at time s ($\mathfrak{S}_s$ stands for the σ -field at time s). We will use notation inspired from conditioning to distinguish between the versions of the forecasts as seen from different time periods. Specifically, $D_{s,t}|\mathfrak{S}_r$, $F_{s,t}|\mathfrak{S}_r$, $\Delta_{s,t}|\mathfrak{S}_r$ and $U_{s,t}|\mathfrak{S}_r$ refer to the random variables corresponding to $d_{s,t}$, $f_{s,t}$, $\delta_{s,t}$ and $u_{s,t}$, as seen from period r . Thus, $F_{s,t}|\mathfrak{S}_r$, $\Delta_{s,t}|\mathfrak{S}_r$ and $U_{s,t}|\mathfrak{S}_r$ are all random for $r < s$, but $f_{s,t} = F_{s,t}|\mathfrak{S}_s$, $\delta_{s,t} = \Delta_{s,t}|\mathfrak{S}_s$ and $u_{s,t} = U_{s,t}|\mathfrak{S}_s$ are deterministic. We clarify the conditioning of forecasts in Appendix B. With this notation equation (3) becomes

$$U_{s,t}|\mathfrak{S}_r = (\Delta_{s,t} - \Delta_{s-1,t})|\mathfrak{S}_r. \quad (4)$$

Following [17], we make the following assumptions on the update random variable:

- (A1) *No Learning*: Nothing is learned about $U_{s,t}$ before period s . $U_{s,t}$ indeed represents the additional information learned in period s . For $r < s$, all $U_{s,t}|\mathfrak{S}_r$ have the same distribution as the generic random variable $U_{s,t} := U_{s,t}|\mathfrak{S}_{s-1}$. *No Learning* implies that $U_{s,t}$ and $U_{r,w}$ are uncorrelated for $s \neq r$.
- (A2) *Stationarity*: $U_{s,t} = U_{s+h,t+h}$ in distribution for any increment h .
- (A3) *Zero Expected Value*: $E(U_{s,t}) = 0$. Appendix D discusses how to proceed when this assumption fails.

We base forecast evolution on perceived ramp age forecasts as opposed to fractional forecasts so that (A2) becomes a reasonable assumption. We observe that fractional updates $(f_{s,t} - f_{s-1,t})$ tend to be smaller when $f_{s,t}$ is close

to either 0 or 1. Thus, the dependence of fractional updates on ramp ages prohibits a stationarity assumption on fractional updates.

The main interest here is in calculating the variance of the forecast updates. Forecasts made for periods too far into the future are not useful, so we have a finite forecast horizon H . Thus, perceived age forecasts $\Delta_{s,t}|\mathfrak{S}_r$ for $s < t - H$ will not be defined. Before period $t - H + 1$, the only information available on period t demand is $\delta_{t-H,t}$, so for $r \leq t - H$, $\Delta_{s,t}|\mathfrak{S}_r = \Delta_{s,t}|\mathfrak{S}_{t-H}$ and $\Delta_{t-H,t}|\mathfrak{S}_r = \delta_{t-H,t}$. In the more interesting case of $s \geq t - H$ and $r \geq t - H$, from Equation (4), the perceived age forecast is

$$\Delta_{s,t}|\mathfrak{S}_r = \delta_{t-H,t} + \sum_{j=t-H+1}^s U_{j,t}|\mathfrak{S}_r = \delta_{t-H,t} + \sum_{j=t-H+1}^{r \wedge s} u_{j,t} + \sum_{j=r+1}^s U_{j,t} \quad s \geq t - H, r \geq t - H \quad (5)$$

where $r \wedge s = \min(r, s)$. The second equality follows from the assumption of *No Learning* about updates $U_{j,t}$ before they are observed. As r increases, the deterministic component of perceived age forecasts (for fixed s and t) grows, and the forecast eventually becomes deterministic at $r = s$ ($\delta_{s,t} = \Delta_{s,t}|\mathfrak{S}_s$). It follows from Equation (5) that the stochastic parts of $\Delta_{s,t}|\mathfrak{S}_r$ and $\Delta_{s+h,t+h}|\mathfrak{S}_{r+h}$ have the same distribution for any increment h .

Obtaining forecasts via Equation (5) has a nice feature. The mean square error of perceived age forecasts is non-increasing and goes to zero as s approaches t for any r , i.e.,

$$E[(\Delta_{t,t} - \Delta_{s,t})^2|\mathfrak{S}_r] = \sum_{j=(s \vee r)+1}^t \text{var}(U_{j,t}) + \left\{ \sum_{j=(s \wedge r)+1}^r u_{j,t} \right\}^2$$

where $s \vee r = \max(s, r)$. The last equality follows from the *No Learning* assumption.

Perceived age forecasts are unbiased if (A3) holds. Unbiasedness indicate that an observation $\delta_{s,t}^{p,tec}$ is roughly equal to the expected value of $\Delta_{t,t}^{p,tec}$ where the expectation is taken relative to information available in period s . Although $\Delta_{s,t}^{p,tec}|\mathfrak{S}_r$ and $\Delta_{s,t}^{p,tec}|\mathfrak{S}_{r+1}$ have different means, perceived age forecasts as given by equation (5) satisfy

$$E(\Delta_{t,t}^{p,tec}|\mathfrak{S}_s) = \Delta_{s,t}^{p,tec}|\mathfrak{S}_s = \delta_{s,t}^{p,tec}. \quad (6)$$

Thus, as a consequence of the *No Learning* and *Zero Expected Value* assumptions, perceived age forecasts are unbiased. When forecasts are made from s for t (where $s \leq t$), we coin the term (forecast) *lag* for $t - s$. We say that $\delta_{s,t}^{p,tec}$ ($d_{s,t}^p$) is a *lag biased forecast* if $E(U_{s,t}^{p,tec})$ ($E(V_{s,t}^p)$) differs from zero by a deterministic function of the lag. In Appendix D, we illustrate how our work can be extended to accommodate lag bias.

Fractional forecasts are related to perceived age forecasts via a nonlinear ramp curve $\hat{R}$ (see Equation (2)). Since perceived age forecasts are unbiased, there will be bias in fractional forecasts $f_{s,t}^{p,tec}$. In regions where the ramp function is approximately linear, this *nonlinearity-induced bias* will be small. We will revisit the magnitude of the bias in the numerical experiments section.

Perceived age forecasts can be constructed as conditional expectations, i.e. we could define $\delta_{s,t} = E(\Delta_{t,t}|\mathfrak{S}_s)$ (see section 2.7. of [21]). If we did, by using the properties of the conditional expectations, (A3) and the uncorrelated updates part of (A1) could be deduced (see Appendix C). These can also be deduced when $\delta_{s,t}$ are minimum mean-squared error forecasts (see [17]). Also note that perceived age forecasts constitute a martingale, i.e. $E(\Delta_{s,t}|\mathfrak{S}_r) = \delta_{r,t}$ if $r \leq s \leq t$.

5 Normalized Updates

In the previous section, we have illustrated the evolution of perceived age forecasts by focusing on a fixed product (p, tec). This allowed us to discuss the stochastic mechanisms with notational simplicity. We will now obtain some statistics on these mechanisms.

The updates, being a function of the time lag and the product, can be calculated as

$$u_{s,t}^{p,tec} = \hat{R}^{-1}(f_{s,t}^{p,tec}) - \hat{R}^{-1}(f_{s-1,t}^{p,tec}). \quad (7)$$

These figures are called normalized updates because their distribution depends on the difference between t and s (by (A2)), but not on t , s or the ramp age.

At every time period, an update vector for a currently ramping technology will be constructed by putting all updates (H updates for each product family) into a vector u_s^{tec} . For example, assume that memory and X86 are the only two product families. At time s , for technologies 10 and 8 and product families memory and X86 chips, we would have vectors u_s^8 and u_s^{10} as in Equation 8. In period $s + 1$, u_s^{10} evolves into u_{s+1}^{10} .

$$\begin{aligned} u_s^8 &= [u_{s,s}^{X86,8}, \dots, u_{s,s+H-1}^{X86,8}, u_{s,s}^{Mem,8}, \dots, u_{s,s+H-1}^{Mem,8}] \\ u_s^{10} &= [u_{s,s}^{X86,10}, \dots, u_{s,s+H-1}^{X86,10}, u_{s,s}^{Mem,10}, \dots, u_{s,s+H-1}^{Mem,10}] \\ u_{s+1}^{10} &= [u_{s+1,s+1}^{X86,10}, \dots, u_{s+1,s+H}^{X86,10}, u_{s+1,s+1}^{Mem,10}, \dots, u_{s+1,s+H}^{Mem,10}] \end{aligned} \quad (8)$$

This specific construction lets us observe several update vectors (one for each active technology) in a single period. Note that different technologies are not put into the same update vector. This is because they are assumed to be independent (see (I2)). In general, u_s^{tec} has entries for each product p and each lag $t - s$, $0 \leq t - s < H$.

For convenience and stationarity (see (A2)), we assume normality of update vectors:

(I3): U_s^{tec} is normally distributed with covariance matrix Σ for all (s, tec) .

Note that $U_{s_1}^{tec_1}$ and $U_{s_2}^{tec_2}$ are stochastically independent provided that $s_1 \neq s_2$ or $tec_1 \neq tec_2$. If $s_1 \neq s_2$, this is a consequence of (A1) and (I3). If $tec_1 \neq tec_2$, it is a consequence of (I2).

Assumptions (A1), (A3), (I2) and (I3) imply that update vectors U_s^{tec} are independently and identically distributed as $N(0, \Sigma)$. Though updates in forecasts for different technologies are independent, the components of the vector U_s^{tec} will be dependent among themselves. Thus, we still capture demand correlations among different product families as well as among time periods.

However, with the above construction, not all components of a given update vector will be observed at all periods. If $tec = 10$ is introduced into family X86 at time t , then no perceived age forecast $\delta_{s,*}^{X86,10}$ will be available in period $s = t - H - 1$. Two periods later in period $t - H + 1$, the update vector u_{t-H+1}^{10} will have a single observed element, $u_{t-H+1,t}^{X86,10}$. In the next period, the update vector u_{t-H+2}^{10} will have two observed elements $u_{t-H+2,t}^{X86,10}$ and $u_{t-H+2,t+1}^{X86,10}$, and so forth. At a given point in time, a particular technology may be used for some product families, but not for the others. In that case, only forecast updates of product families using that technology will be observed. Basically, some data will be missing in many (or most) of the update vectors. Conceptually, this is not a big issue; nevertheless it poses numerical obstacles for the estimation of the covariance matrix Σ .

6 Covariance Matrix Estimation

As we mentioned before, evolution model for family forecasts is similar to perceived age forecasts with $d_{s,t}^p$ and v_s playing the roles of $\delta_{s,t}^{p,tec}$ and u_s^{tec} . Note that assumptions analogous to (A1)-(A3) and (I3) must be made to deduce that family forecast update vectors, $v_s = [v_{s,s}^{X86}, \dots, v_{s,s+H-1}^{X86}, v_{s,s}^{Mem}, \dots, v_{s,s}^{PPC}, \dots]$ are normally distributed with $N(0, \Lambda)$. Then, Λ can be estimated using classical techniques because in this case update vectors have no missing elements.

We will now discuss the estimation of Σ , the covariance matrix for perceived age updates. Denote the length of an update vector u_s^{tec} with $P \cdot H$, the number of product families P times the length of the forecast horizon H . Then, u_s^{tec} is an observation from a normal random vector with the $(P \cdot H) \times (P \cdot H)$ covariance matrix Σ . Note that the components of u_s^{tec} are indexed by (product family) $\times$ (time lag), and so are the entries of Σ .

Estimation will be based on a maximum likelihood framework. Suppose that the vectors u_s^{tec} are numbered from 1 to N , to obtain the sample $\{u_i : i = 1..N\}$. N is approximately the number of time periods times the average number of active technologies being produced at a given time. The MLE estimator $\hat{\Sigma}$ of Σ would solve the following minimization problem:

$$\min_{\Sigma} \frac{N}{2} \log |\Sigma| + \frac{1}{2} \sum_{i=1}^N u_i \Sigma^{-1} u_i^T.$$

The solution to this minimization problem is easily found when no data is missing (see Anderson [22]). When some data is missing, an iterative procedure, called the EM algorithm maximizing the likelihood function given the observed data, is available. The details of the EM algorithm are found in Schafer [23].

7 Demand Covariances

Suppose that we are in period r , $r < t$ where t denotes future periods. Then, $d_{r,t}^{p,tec}$ are current forecasts and $D_{t,t}^{p,tec}|\mathfrak{S}_r$ are true demands as seen from period r . We want to estimate covariances of $D_{t,t}^{p,tec}|\mathfrak{S}_r$. The procedure that estimates R , Λ , Σ and calculates demand covariances will be called SeDFAM (acronym for **S**emiconductor **D**emand **F**orecast **A**ccuracy **M**odel). Table 1 summarizes the computations required by SeDFAM. So far we have discussed the first seven steps.

Inputs : Demand forecasts, $d_{s,t}^{p,tec}$
1. Compute family forecasts $d_{s,t}^p$ and fractional forecasts $f_{s,t}^{p,tec}$
2. Fit ramp function $\hat{R}$ to fractional forecasts
3. Compute perceived age forecasts $\delta_{s,t}^{p,tec} = \hat{R}^{-1}(f_{s,t}^{p,tec})$
4. Compute family forecast update vectors v_s (analogous to u_s^{tec}) from $d_{s,t}^p$
5. Estimate family forecast update covariance matrix, $\hat{\Lambda}$ (analogous to $\hat{\Sigma}$)
6. Compute perceived age forecast update vectors u_s^{tec} from $\delta_{s,t}^{p,tec}$
7. Estimate perceived age forecast update covariance matrix, $\hat{\Sigma}$ (using the EM algorithm)
8. Use $\hat{R}$, $\hat{\Lambda}$, $\hat{\Sigma}$ to compute variances and covariances of demands as seen in period r

Table 1: Steps of the SeDFAM

We now discuss recovering the covariances of technology demands (Step 8 in Table 1). By the definition of fractional demands in Equation 1, we obtain:

$$D_{t,t}^{p,tec}|\mathfrak{S}_r = \{D_{t,t}^p|\mathfrak{S}_r\} \{(F_{t,t}^{p,tec}|\mathfrak{S}_r) - (F_{t,t}^{p,tec+}|\mathfrak{S}_r)\}$$

where tec stands for a technology and $tec+$ denotes the next technology introduced after tec . We suppress $\mathfrak{S}_r$ in the notation for brevity. From the independence of product family and fractional demands, we obtain

$$\begin{aligned} Cov(D_{t_1,t_1}^{p_1,tec_1}, D_{t_2,t_2}^{p_2,tec_2}) &= Cov(F_{t_1,t_1}^{p_1,tec_1} - F_{t_1,t_1}^{p_1,tec_1+}, F_{t_2,t_2}^{p_2,tec_2} - F_{t_2,t_2}^{p_2,tec_2+}) E(D_{t_1,t_1}^{p_1} D_{t_2,t_2}^{p_2}) \\ &+ Cov(D_{t_1,t_1}^{p_1}, D_{t_2,t_2}^{p_2}) E(F_{t_1,t_1}^{p_1,tec_1} - F_{t_1,t_1}^{p_1,tec_1+}) E(F_{t_2,t_2}^{p_2,tec_2} - F_{t_2,t_2}^{p_2,tec_2+}). \end{aligned} \quad (9)$$

$Cov(F_{t_1,t_1}^{p_1,tec_1} - F_{t_1,t_1}^{p_1,tec_1+}, F_{t_2,t_2}^{p_2,tec_2} - F_{t_2,t_2}^{p_2,tec_2+})$ is the most interesting term. This covariance is zero, if tec_2 comes after tec_1+ . Suppose $tec = tec_2 = tec_1$. Using the independence of different technologies,

$$Cov(F_{t_1,t_1}^{p_1,tec} - F_{t_1,t_1}^{p_1,tec+}, F_{t_2,t_2}^{p_2,tec} - F_{t_2,t_2}^{p_2,tec+}) = Cov(R(\Delta_{t_1,t_1}^{p_1,tec}), R(\Delta_{t_2,t_2}^{p_2,tec})) + Cov(R(\Delta_{t_1,t_1}^{p_1,tec+}), R(\Delta_{t_2,t_2}^{p_2,tec+})). \quad (10)$$

Note that the vector

$$[\Delta_{t_1,t_1}^{p_1,tec} \Delta_{t_2,t_2}^{p_2,tec} \Delta_{t_1,t_1}^{p_1,tec+} \Delta_{t_2,t_2}^{p_2,tec+}] \quad (11)$$

is normally distributed with the expected value $[\delta_{r,t_1}^{p_1,tec} \delta_{r,t_2}^{p_2,tec} \delta_{r,t_1}^{p_1,tec+} \delta_{r,t_2}^{p_2,tec+}]$. To get the covariance, let $\hat{\Sigma}(p_1, p_2, i, j)$ be the entry in i th row and j th column of submatrix of $\hat{\Sigma}$ that corresponds to product families p_1 and p_2 . Pairs

of the random variables in (11) have zero covariance if they belong to different technologies. Otherwise, suppose without loss of generality $t_1 \leq t_2$. Then we estimate $cov(\Delta_{t_1, t_1}^{p_1, tec}, \Delta_{t_2, t_2}^{p_2, tec})$ by

$$\sum_{i=1}^{i=t_1-r} \Sigma(p_1, p_2, i, i + t_2 - t_1).$$

Given the covariance matrix $\hat{\Sigma}$ and means of the vector in (11), computing the covariance in (10) requires 2-dimensional numerical integrations. Such integrations become very tedious when $\hat{R}$ is a piece-wise quadratic spline function, as in our numerical experiments. Thus, we will calculate the expectations and the covariances involving $F_{t,t}^{p,tec}$ in Equation (9) using a Monte-Carlo approach. The covariances and expectations involving $D_{t,t}^p$ are easily obtained from $\hat{\Lambda}$ and the observed data. This completes our definition of SeDFAM.

8 Studying a Base Case with SeDFAM

In this section, we will study the effectiveness of SeDFAM with the simulated forecasts. We simulate six product families with several ramps and study the covariances between two of the families (See Figure 2).

Figure 2: Two product family demands during 60 months

8.1 Simulating a Forecast History

We briefly describe the algorithm (see Figure 3) used to randomly generate forecast history data $d_{s,t}^{p,tec}$, for $t \leq now$ where *now* stands for the current time period. Family forecasts $d_{s,t}^p$ are generated using a given update covariance matrix Λ_0 . We use equation (5) (replacing Δ with D and U with V) to generate family forecasts starting from forecasts made $H = 6$ periods in advance.

We use a given perceived ramp age update covariance matrix Σ_0 to generate age forecasts, $\delta_{s,t}^{p,tec}$, through Equation (5). Fractional ramp forecasts, $f_{s,t}^{p,tec}$ are obtained from ramp age forecasts via equation (2). The ramp curve we

Inputs: R_0 , Λ_0 , Σ_0

Figure 3: Flowchart for Simulating Forecasts

use is also given and is a symmetric cubic polynomial: $R_0(\delta) = 3\delta^2 - 2\delta^3$. Sometimes simulated fractional forecasts that are made in the same period are out-of-order, i.e. $f_{s,t}^{p,tec} > f_{s,t+1}^{p,tec}$. Since newer technologies replace older ones, we assume the $f_{s,t}^{p,tec}$ are nondecreasing in t for fixed s . To achieve this, out-of-order fractional forecasts are sorted. Lastly, product demand forecasts are obtained from the following equation (see Figure 3):

$$d_{s,t}^{p,tec} = \{d_{s,t}^p\} \{f_{s,t}^{p,tec} - f_{s,t}^{p,tec+}\}.$$

8.2 Two Heuristics: Allocation and Proportion Schemes

In order to measure the effectiveness of SeDFAM against other approaches that might be attempted in practice, we have considered two other schemes as well. In the first scheme, family lag- h forecast error variances $(\sigma_h^p)^2 = \text{var}(D_{t,t}^p - D_{t-h,t}^p | \mathfrak{I}_{t-h})$ are calculated from the historical family demand. In each period t ($t > \text{now}$), they are *allocated* to products by the fractional forecasts made in period *now*: $\text{var}((D_{t,t}^{p,tec} - D_{now,t}^{p,tec}) | \mathfrak{I}_{now}) = (d_{now,t}^{p,tec})(\sigma_h^p)^2$ where $t = \text{now} + h$. In this *Allocation Scheme*, product demands are treated as if they were independent.

In the second scheme, called the *Proportion Scheme*, assume that fractional forecast errors in product demands depend only on the lag, and is independent of product family and technology. We construct a new random variable Ψ_{t-s} and assume that

$$\psi_{t-s} \sim \frac{d_{s,t}^{p,tec} - d_{t,t}^{p,tec}}{d_{s,t}^{p,tec}} \text{ for } t - H \leq s < t. \quad (12)$$

We call ψ_{t-s} the *proportional lag update*. Then, in period *now*,

$$\text{var}((D_{t,t}^{p,tec} - D_{now,t}^{p,tec}) | \mathfrak{I}_{now}) = (d_{now,t}^{p,tec})^2 \text{var}(\Psi_{t-now}) \quad (13)$$

8.3 Base Case

We have structured our numerical study around a base case. In coming up with the base case, we have relied on our interviews with people in the semiconductor industry (see [1]). We set the available forecast history to 60 months of data, and the forecast horizon H to six months. We calculated a product family update covariance matrix using

data obtained from a semiconductor manufacturer and based our “true” Λ_0 on that. In the base case, on average, every 8-10 months a new technology is introduced. Technologies stay active almost 24 months. Also note that one of the product family demands has a linear trend of going up whereas the other is stable (see Figure 2).

Our experimental setup is composed of 10 replications. Replications have *now* dates that are two months apart. They cover the start, middle and end of the ramp, so that SeDFAM can be evaluated at different phases of the ramp. Each replication uses 60 months of forecast history.

In applying SeDFAM to the base case, we follow the steps outlined in Table 1. We comment on step 2. It is possible to estimate a different ramp curve for each technology or for each product family. However, if a ramp curve is estimated from only 4-5 ramps, it can not be estimated very accurately. Since the forecast history, in practice, often does not go beyond 4-5 ramps, we suggest that a single $\hat{R}$ be fit to historical data $f_{s,t}^{p,tec}$ from all families and all technologies, using 48-60 periods. We fit a piecewise quadratic spline, $\hat{R}$, to the fractional forecast data. The spline has three knots with $\hat{R}(0) = 0$, $\hat{R}(L) = 1$. It is constrained to have vanishing derivatives at the endpoints $\{0, L\}$.

Estimation of Λ is a straightforward application of the Heath-Jackson scheme ([17]). The estimated $\hat{\Lambda}$ can be directly compared to Λ_0 . This direct comparison is not as meaningful for Σ , because it is influenced by our sorting of fractional forecasts. Moreover, practically speaking the quality of the estimates of Λ and Σ are not as important as estimates of the covariances of forecast errors.

We are especially interested in errors in forecasted capacity requirements for specific tools. We focus on a critical tool, called Ctool, that has processing times (per job) of 1 hour for (A, tec) , 1.3 hours for $(A, tec+)$, 0.7 hours for (B, tec) and 1 hour for $(B, tec+)$. The technology (*tec*) is introduced on product family A in the 61st month, and its successor (*tec+*) in the 68th month. Those technologies (*tec* and *tec+*) are introduced on product family B in the 64th and the 73rd months. The Ctool is not used at all before month 61. Let $C_{now,t}$ be the capacity demand for Ctool in period t as seen from *now*. Capacity demands for Ctool are obtained by multiplying product demand forecasts by processing times and summing. Let $C_{now} = [C_{now+1,now+1}, C_{now+2,now+2}, \dots, C_{now+H,now+H}]$.

Starting from a single forecast history covering all periods t , $t \leq now$, we randomly generate a sample of 5000 independent future product demands ($D_{t,t}^{p,tec}$), $t > now$. “True” values of all performance measures are derived from these future product demands.

8.4 SeDFAM vs. Heuristics: Estimation and Decision Making

We apply SeDFAM, and the Allocation and Proportion heuristics to generate estimated variances for the base case. For each lag, we average the following measure over all replications, and call it the *fractional error in variance*

$$\frac{\text{Estimated Variance} - \text{True Variance}}{\text{True Variance}}$$

Figure 4 depicts the relative performance of the heuristics against SeDFAM, in predicting the demand variance for Ctool.

For the base case, it appears that the Allocation Scheme underestimates the variances: Its estimates are uniformly 70% of the true variances. Demands for succeeding technologies of the same family are negatively correlated. The *Allocation Scheme* does not capture these correlations, and that results in underestimation of variances. On the other hand, the *Proportion Scheme* overestimates variances by 50% to 110%. Large-valued $D_{s,t}^{p,tec}$ lead to overly large variance estimates (see Equation (13)).

We now want to see the business implications of inaccurate variance estimation using a simple capacity acquisition model. We study six type of tools whose installment lead times vary from 2 months to 12 months. All of these tools have the same processing times as Ctool, described in the previous section. Using variance estimates from SeDFAM, tools are bought to satisfy the true capacity demand with a probability of 84.1%. Table 2 depicts actual probabilities of meeting the true capacity demand when capacities are selected according to SeDFAM variance estimates. Similar computations were done for the *Allocation and Proportion Schemes*. For example, for tools with a 2 month lead time (LT=2), demand is met 76.6% of the time when capacity acquisition is based on Allocation estimates, whereas the target is 84.1%. The *Proportion Scheme* overestimates variance and sets capacity levels high and achieves infrequent shortages.

Figure 4: SeDFAM and heuristics' performance in terms of fractional errors in variance of demand for Ctool.

Method	LT=2	LT=4	LT=6	LT=8	LT=10	LT=12	Average error
SeDFAM	83.2 %	82.6 %	83.0 %	83.0 %	83.5 %	83.9 %	0.97%
Allocation	76.6 %	78.1 %	78.5 %	79.1 %	79.7 %	80.0 %	5.52%
Proportion	86.2 %	85.4 %	88.2 %	87.2 %	86.8 %	88.2 %	3.29%

Table 2: Probabilities of meeting capacity demands for tools of lead time=2..12

Since infrequent shortages are achieved at the expense of buying extra capacity, we also report the ratio of expected excess tool capacity to expected capacity required (see Table 3). Ratios are converted to percentages for readability. “True” is the expected excess capacity actually required to meet the demand with a probability of 84.1%. As expected, *Proportion* (*Allocation*) consistently installs too much (not enough) capacity.

Method	LT=2	LT=4	LT=6	LT=8	LT=10	LT=12	Average error
True	22.9 %	34.4 %	33.6 %	35.5 %	33.9 %	33.1 %	-
SeDFAM	22.1 %	32.8 %	33.0 %	34.5 %	33.4 %	32.8 %	1.1%
Allocation	18.8 %	29.1 %	28.5 %	30.5 %	29.5 %	29.1 %	5.4%
Proportion	32.8 %	38.5 %	40.7 %	40.7 %	37.8 %	38.5 %	11.4%

Table 3: Ratio of excess capacity to expected capacity required.

Second, we study the effects of inaccurate estimation on revenue prediction. We now suppose that prices are \$1 per unit of (A, tec) , \$1.3 for $(A, tec+)$, \$0.7 for (B, tec) and \$1 for $(B, tec+)$. We assume that all demand is satisfied and estimate the variance of the total revenue over the next six months. The difference between the SeDFAM variance estimate and the true variance is scaled by the true variance to obtain fractional errors in 6-month revenue variances. Fractional errors in variances are also calculated for the heuristics. See Figure 5. In the base case, demands are positively correlated in time. The *Allocation Scheme* does not capture these correlations, and drastically underestimates the 6 month variance. *Proportion Scheme* estimates are small (large) when the demand

forecasts $D_{now,t}^{p,tec}$ are small (large) (see Equation (13)).

Figure 5: Fractional errors in 6 month revenue variance estimates.

8.5 Nonlinearity Bias

Near the end of section 4, we discussed nonlinearity bias in $f_{s,t}^{p,tec}$. There are three interesting quantities to compare. First, the fractional simulated forecast $f_{now,t}^{p,tec}$, which comes from the historical data. Second, the true mean of the fractional demand, $E(F_{t,t}^{p,tec}|\mathcal{S}_{now})$. Third, SeDFAM creates $\hat{R}$ and $\hat{\Sigma}$. We use $\hat{\Sigma}$ to estimate the distribution of $(\Delta_{t,t}^{p,tec}|\mathcal{S}_{now})$, and use $\hat{R}$ to estimate $E(\hat{R}(\Delta_{t,t}^{p,tec}|\mathcal{S}_{now}))$. All three quantities are converted from estimates of fractional demands to estimates of demand for Ctool capacity, resulting in the “Forecast”, the “True Estimate”, and the “SeDFAM Estimate”.

The bias percentage is the difference between either the Forecast or the SeDFAM estimate and the True Estimate, divided by the Forecast (see Figure 6). This figure depicts nonlinearity bias percentage in 6-month out forecasts for each of the ten replications whose *now* dates range from the 60th to the 78th month. Towards the 78th month, the *tec* ramp is about to end, so nonlinearity bias is almost nonexistent. The nonlinearity bias is too small to have a significant effect on our results.

9 Robustness of SeDFAM

In this section, we test the robustness of SeDFAM against variations in the base case parameters: forecast history, ramp variability, length of ramp lives, skewedness of ramp curves, covariance structures and forecast horizon. SeDFAM’s performance is measured in terms of its accuracy in predicting capacity demand covariance matrices.

Step 8 of Table 1 computes the variances and the covariances of forecasts made in period *now*. These data are used to generate an estimate $\hat{\Gamma}$ of the covariance matrix of the vector C_{now} of capacity demands for Ctool. Let Γ be the true capacity demand covariance matrix. The principal measure for the performance of SeDFAM is $F(\Gamma)$, defined as

Figure 6: Nonlinearity bias in 6 months in advance forecasts.

$$F(\Gamma) = \frac{\|\hat{\Gamma} - \Gamma\|_F}{\|\Gamma\|_F}$$

where subscript F denotes the Frobenious norm. For 10 replications of the Base Case, $F(\Gamma)$ has an average value of 8.5%.

9.1 Experiment on Forecast History

In the base case, there are 60 months of forecast history. Naturally, with more historical data, we can estimate Γ better (see figure 7). Note that estimation errors are within 15% both with history of 60 and 90 months. Thus, by examining forecast histories beyond 60 months, covariance estimates can not be improved very much. On the other hand, errors go up to almost 20% when forecast history is halved down to 30 months. As a result, we conclude that 45-60 months of forecast history will deliver satisfactory covariance estimates.

9.2 Experiment on Ramp Variability

We test the effectiveness of our EM method when the variability in the ramp forecasts changes. For that, we multiply Σ_0 by 2.25 and by 0.25 to obtain two versions that differ from the base case only by ramp variability. We depict the $F(\Gamma)$ measure (see Table 4) and conclude that the effectiveness of SeDFAM does not depend on ramp variability.

9.3 Experiment on the Length of Ramp Lives

Ramp life is the length of the S-curve in Figure 1 and should not be confused with the product lives. It is the time from the introduction of a technology until the obsolescence of the previous technology. Ramp life averages 14 months in the base case. We experiment with average ramp lives of 10 and 18 months (see Table 4). The improvement in performance is probably due to the increase in the number of observed data elements per month. The impact on SeDFAM's accuracy is small.

Figure 7: Effectiveness of SeDFAM estimates as forecast history varies.

9.4 Experiment on Skewedness of Ramp Curves

In this section, we investigated whether the skewedness of the ramp curve has any effect on the effectiveness of our method. For that, we used a left-skewed, right-skewed and a symmetric curve as ramp curves. A symmetric ramp curve satisfies:

$$R_0(t) + R_0(L - t) = 1 \text{ for } 0 < t < L.$$

The “=” above becomes “>” (“<”) for a left-skewed (right-skewed) ramp curve. Table 4 shows skewedness has a weak effect on performance.

9.5 Experiment on Covariance Structures

Both Λ and Σ contain covariances across time and between product families. In this experiment, we first test the response of our method to higher or lower correlations across time and in product family demands while everything else is kept constant. We regulate the time-wise covariances by scaling the appropriate submatrices of Λ . High (low) correlation in the family demand section of Table 4 refers to a situation where month to month demand covariances inside a product family are approximately increased (decreased) by 100% (50%) and covariances across families are approximately increased (decreased) by 50% (30%). As Table 4 indicates, SeDFAM effectiveness does not change with the amount of month-to-month correlations. Second, above experiment is repeated with the ramp age covariance matrix, Σ . From Table 4, we conclude that relative magnitude of both ramp age and family covariances have a small effect on SeDFAM performance.

9.6 Experiment on Forecast Horizon

Lastly, we measure the effectiveness of our scheme under different forecast horizons, $H = 6, 8, 10$ or 12 . Table 4 shows the effect of longer forecast horizons on $F(T)$. It is hard to detect a consistent trend from this table. Thus, we conclude that forecast horizon does not have a clear effect on performance.

Months	60	62	64	66	68	70	72	74	76	78	Average
Base case	7.3	15.6	6.0	15.4	6.8	6.5	4.5	10.8	5.9	6.1	8.5
Ramp variability											
Less	7.1	16.9	9.5	20.4	6.9	7.1	4.6	11.2	6.0	6.1	9.6
More	7.2	16.3	7.6	13.6	4.5	6.5	3.8	11.1	5.9	6.2	8.3
Length of Ramps											
L=10	12.1	11.5	10.7	9.4	3.4	13.1	12.3	12.8	8.0	9.2	10.2
L=18	2.4	6.7	3.5	6.0	5.7	12.9	9.7	10.4	8.4	7.0	7.3
Ramp skewedness											
Left	9.5	17.5	5.4	19.4	8.5	6.2	6.2	12.2	6.7	6.1	9.8
Right	5.7	23.3	6.7	20.0	8.8	7.2	5.5	12.5	6.9	6.2	10.3
Serial correlation in family demand											
Low	8.5	16.3	6.3	13.3	6.1	6.2	4.6	10.8	5.8	6.2	8.4
High	5.9	15.0	7.9	17.0	7.3	6.8	4.4	10.7	6.1	6.1	8.7
Serial correlation in perceived ages											
Low	6.2	17.0	7.1	18.8	5.2	7.4	4.8	11.0	6.1	6.1	9.0
High	8.4	16.5	7.1	15.5	5.4	6.5	4.0	10.9	6.0	6.2	8.7
Forecast Horizon, H											
$H = 8$	7.8	11.9	6.7	15.3	17.6	21.3	29.3	30.3	15.1	29.4	18.5
$H = 10$	12.1	16.4	8.4	12.5	10.4	22.2	8.8	6.6	21.1	11.4	13.0
$H = 12$	11.5	19.1	18.0	34.1	11.7	19.9	8.2	5.3	18.0	5.5	15.1

Table 4: Robustness of SeDFAM in terms of $F(\Gamma)$

10 Industry Example

A semiconductor manufacturer provided us with an industrial data set of annual forecasts with quarterly time buckets, and with $H = 10$ quarters. We have 5 years of data from 1994 to 1998. We studied 4 product families and 5 technologies. However, not all technologies are used on all 4 families in all time periods. Figure 8 depicts $d_{s,t}^p$ and $f_{s,t}^{p,tec}$. A single curve in Figure 8 represents either $d_{s,t}^p$ or $f_{s,t}^{p,tec}$ as t ranges from s to $s + H$.

10.1 Validating SeDFAM Assumptions

The first step in applying SeDFAM is checking the validity of the assumptions. SeDFAM is built on two critical independence assumptions, (I1) and (I2). (I1) implies the independence of the perceived age update u and the family update v . Hence, whether a technology is delayed or expedited has no effect on family demands. To improve significance of the statistical tests, we pool data from all families together. Then, for each forecast lag (6 lags), we test (for (I1)) the componentwise independence of the family and perceived age update vectors that are observed in the same year. Thus, we have 6 tests, each test has on the average 20 sample points depending on the pace of the ramps. Assumption (I2) implies the independence of two perceived age updates calculated from two *sequential technologies* (one coming directly after another) on the same family. In other words, delaying one technology should not significantly delay the next one. For each forecast lag (6 lags), we test (for (I2)) the componentwise independence of two *sequential technologies'* perceived age updates observed in the same year. There are again 6 tests, and each have 11-18 sample points.

Independence is tested with a likelihood ratio based parametric test by assuming normality of update vectors (see pp 220 Bickel and Doksum [24]). The smallest p-value of the 12 tests (6 for (I1) and 6 for (I2)) is 0.29. Consequently, (I1) and (I2) assumptions are validated for the industrial forecasts.

Figure 8: Family forecasts, $d_{s,t}^p$ and fractional ramp forecasts, $f_{s,t}^{p,tec}$. Both are for four product families A, B, C and D. Each style of curve in the fractional ramp forecasts represents a different technology.

Assumption (A1) cannot be tested fully with the given forecast data because, in year r , there is no information about a forecast made in year s ($s > r$). Instead, we test the last assertion of (A1). It requires that updates computed in different periods be uncorrelated (or independent for Normal updates). We tested independence of family (perceived age) updates with 12 (14) sample points. There is no indication of dependence in family updates and the test of perceived age update has a p-value of 0.43. Consequently, we conclude that updates are uncorrelated.

Assumption (A2) requires that updates be stationary. In figure 9, we plot perceived age updates versus ramp ages to see if updates have a pattern or trend. There is no significant trend so assumption (A2) seems reasonable.

Figure 9: Perceived age updates vs. ramp ages

Assumption (A3) requires that family and perceived age updates (each of length 2) have mean zero. In this test, sample size is 16 for family updates and between 19 and 21 for perceived age updates. Our tests find only two-year-out family updates unbiased. One-year-out family updates have negative bias. In other words, family forecasts are

optimistic initially and they are decreased to realistic levels one year before the demand is observed. Our tests find perceived age forecasts positively biased, i.e., initially ramp schedules are overly pessimistic. In summary, forecast data indicates biased family updates and perceived age updates. See Appendix D for discussion on how to adapt SeDFAM for these “Lag Biases” in updates.

10.2 Effectiveness of SeDFAM for Industrial Forecasts

After justifying SeDFAM assumptions in the previous section, we use the industrial data to test how SeDFAM responds to lower update frequencies. We consider three cases differentiated by forecast update frequency and forecast time buckets: (Annual,Quarterly), (Semiannual,Quarterly) and (Quarterly,Quarterly). We estimate Λ_0 and Σ_0 from the industrial data for families A and B of Figure 8 and compute $F(\Gamma)$ as in Table 4. In all cases, we run SeDFAM with $H = 8$ quarters and 20 quarters of forecast history. In the (Annual, Quarterly) case, in quarter t we see demand forecasts for quarters $t, \dots, t + 8$. The most recent previous forecast was made in quarter $t - 4$, for quarters $t - 4, \dots, t + 4$. Those forecasts have 5 overlapping quarters $t, \dots, t + 4$ and two families, so Λ and Σ are 10 x 10 matrices.

We summarize the performance of SeDFAM in Table 5. The second and third columns contain the approximate number of observations used in estimating the Λ, Σ matrices respectively. As forecast update frequency increases (from once to four times in a year), the size of the covariance matrices and the number of observations both grow. In the (Annual,Quarterly) and (Semiannual,Quarterly) cases, sample sizes are small with respect to the size of Σ so the EM algorithm diverges without yielding a $\hat{\Sigma}$. To circumvent this, we use a Bayesian version of SeDFAM (first three rows of Table 5). The last row shows the performance of (the Frequentist version) SeDFAM and is included for comparison purposes.

Three Bayesian and One Frequentist Cases	Sample size for Λ	Sample size for Σ	Size of Λ, Σ	$F(\Gamma)$ by Quarters						Average $F(\Gamma)$
				30	32	34	36	38	40	
Annual,Quarterly,Bayes.	5	10	10x10	63%	51%	35%	37%	40%	42%	44.7%
Semiannual,Quarterly,Bayes.	10	22	14x14	44%	39%	28%	27%	29%	38%	34.2%
Quarterly,Quarterly,Bayes.	20	50	16x16	44%	32%	32%	29%	30%	27%	32.3%
Quarterly,Quarterly,Freq.	20	50	16x16	21%	18%	15%	14%	11%	13%	15.3%

Table 5: SeDFAM effectiveness with industrial forecasts measured in terms of $F(\Gamma)$

Our Bayesian estimation framework uses the Inverted-Wishart prior (see pp. 150 of Schafer [23]). With this prior, we set the expected value of the updates to zero. The expected value of Σ is set equal to a diagonal matrix. All diagonal elements for a given family are equal, meaning that in the prior, variance of the forecast errors is a linearly decreasing function of the forecast lag.

Based on the last two rows of Table 5, we observe that SeDFAM outperforms Bayesian SeDFAM. In the Bayesian framework, our choice of prior adversely affected the quality of the estimation. In summary, we suggest that SeDFAM be used as long as it converges. When it does not converge, use Bayesian SeDFAM preferably with sample-independent priors. Table 5 also indicates that increasing forecast update frequency (going from Annual to Quarterly) helps Bayesian SeDFAM to perform better. This can be explained with the increase in sample sizes.

11 Conclusion

Our results can be used to quantify the risks associated with a given tool purchase plan. Tool purchase decisions are made in the face of a large amount of uncertainty and involve huge investments. Quantification of the risks is closely related to the quantification of forecast accuracy. One of the contributions of this paper is explicitly studying accuracy in the forecasts. It determines how quickly the uncertainty in the forecast of a given month’s demand is

resolved as that month is approached. This helps in specifying the correct amount of flexibility that needs to be built into business strategies.

When a decision is based on inaccurate forecasts, it will be risky. In that case, decision makers may delay the decision to obtain more accurate forecasts. On the other hand, delaying actions creates its own set of risks. Therefore, there is a clear trade off between “postpone” and “commit” decisions. In order to assess the value of the “postpone” option, a characterization of forecasts at the end of the postponement period is necessary. SeDFAM links future forecasts to current forecasts by studying forecast evolution. It captures improvement in forecasts as time goes by. Consequently, SeDFAM is a natural tool to use in postponement vs. commitment trade offs.

Measuring forecast accuracy methodologically (with the covariance matrices) helps monitor forecast quality. By monitoring forecast quality, one can signal when forecasts deteriorate. With SeDFAM, one can even identify whether family demand forecasting or ramp age forecasting is causing the deterioration. Our method can also be easily used in simulating forecasts realistically, because it recognizes dependences among product demands. Such forecast simulations often constitute the primary input for simulation (scenario) based decision making techniques.

By studying the performance of SeDFAM under parametrically varied situations, we have empirically shown that SeDFAM is robust against ramp variability, ramp skewedness and the relative magnitude of time-wise or family-wise covariances. SeDFAM is also robust against the length of ramp lives and forecast horizons. On the other hand, length of forecast history affects performance, especially when forecast history is shorter than 45 months. In its current form, for SeDFAM to work with 5 years of historical data, forecasts should be updated quarterly. If forecast updates are less frequent, a Bayesian version of SeDFAM would be more appropriate.

When our assumptions hold, SeDFAM performs very well. Those assumptions are based on our interviews with semiconductor manufacturers and have been validated using an industrial data set. It is possible to relax some of our assumptions at the expense of added complexity or to simplify our approach in specific situations. However, as it is now, we believe that SeDFAM strikes a good balance between complexity and utility.

Appendix A: Demands with Trend and Seasonality

Assume that our family demands, $\bar{d}_{s,t}^p$, have trend and/or seasonality. Assume that $d_{s,t}^p$ has the same variability structure as $\bar{d}_{s,t}^p$, but without a trend or seasonality effect, i.e.,

$$\bar{d}_{s,t}^p = d_{s,t}^p + tr_t \quad (14)$$

where tr_t is a deterministic function representing trend and seasonality effects at time t .

Note that the update is defined as

$$v_{s,t}^p := \bar{d}_{s,t}^p - \bar{d}_{s-1,t}^p = d_{s,t}^p - d_{s-1,t}^p$$

As a result, neither trend nor seasonality affects the updates or their covariance matrix. Consequently if Equation (14) holds, the key estimation steps are unaffected, and the required changes to SeDFAM are straightforward.

If the trend or seasonality affects variability, $\bar{d}_{s,t}^p$ can not be decomposed as above. In that case, forecast variances must be detrended and/or deseasonalized before forecast evolution (see [17]).

Appendix B: Successive Conditioning

Let us focus on a single product p, tec and drop indices p, tec . We are exploring the relationship between the forecasts $\Delta_{s,t}|\mathfrak{F}_r$ from s for t as seen in r and the forecasts $\Delta_{r,t}|\mathfrak{F}_r$ from r for t ($r < s < t$). We will concentrate only on perceived age forecasts but our discussion is valid for other forecasts, including $f_{s,t}^{p,tec}$ and $d_{s,t}^p$. By equation (6) $\delta_{r,t} = E(\Delta_{t,t}|\mathfrak{F}_r)$.

When $r < s$ and $\mathfrak{F}$ is a filtration (i.e. $\mathfrak{F}_r \subseteq \mathfrak{F}_s$), two interesting comparisons for doubly conditioned random variables are of particular interest:

Lemma : $E(E(\Delta_{t,t}|\mathfrak{F}_s)|\mathfrak{F}_r) = E(\Delta_{t,t}|\mathfrak{F}_r)$, i.e., $E(\Delta_{s,t}|\mathfrak{F}_r) = \delta_{r,t}$. (For proof see theorem 34.4. in [25]).

Observation : $E(\Delta_{t,t}|\mathfrak{F}_s)|\mathfrak{F}_r \neq E(\Delta_{t,t}|\mathfrak{F}_r)$ in distribution, i.e., $\Delta_{s,t}|\mathfrak{F}_r \neq \delta_{r,t}$ in distribution.

For every period t , let $\{\mathcal{F}_s\}$ be a partition of Ω such that $\{\mathcal{F}_s\}$ generates the σ -field $\mathfrak{F}_s$. Each $\mathcal{F}_s$ corresponds to a distinct history up to time s .

We consider a specific example. Consider the Lebesgue measure on the probability space $\Omega = [0, 1]$. Let $\{\mathcal{F}_r\} = \{\{[0, 1]\}\}$ and $\{\mathcal{F}_s\} = \{\{[0, 0.5], (0.5, 1]\}\}$. Then the corresponding σ -fields are $\mathfrak{F}_r = \{\emptyset, [0, 1]\}$ and $\mathfrak{F}_s = \{\emptyset, [0, 1], [0, 0.5], (0.5, 1]\}$. For every ω in the probability space Ω , let $\Delta_{t,t}(\omega) = \omega$, i.e., $\Delta_{t,t}$ is uniformly distributed over $[0, 1]$.

Then, $\delta_{r,t} = E(\Delta_{t,t}|\mathfrak{F}_r) = E(\Delta_{t,t}|[0, 1]) = E(\Delta_{t,t}|\omega \in [0, 1]) = 0.5$.

On the other hand, $(\Delta_{t,t}|\mathfrak{F}_s)$ is a collection of two distinct uniform random variables: $(\Delta_{t,t}|[0, 0.5]) \sim U(0, 0.5)$ and $(\Delta_{t,t}|(0.5, 1]) \sim U(0.5, 1)$. In general, $(\Delta_{t,t}|\mathfrak{F}_s)$ is a collection of random variables, one for each element of $\{\mathcal{F}_s\}$, i.e., for each history observable up to time s . $E(\Delta_{t,t}|[0, 0.5]) = E(\Delta_{t,t}|\omega \in [0, 0.5]) = 0.25$ and $E(\Delta_{t,t}|(0.5, 1]) = E(\Delta_{t,t}|\omega \in [0.5, 1]) = 0.75$. Generally, $\delta_{s,t} = E(\Delta_{t,t}|\mathfrak{F}_s)$ assigns a real number to each period- s history $\mathcal{F}_s$. Finally, $\Delta_{s,t}|\mathfrak{F}_r = (E(\Delta_{t,t}|\mathfrak{F}_s))|\mathfrak{F}_r$ is a random variable. It is equal to 0.25 if $\omega \leq 0.5$, and it is equal to 0.75 if $\omega > 0.5$.

Summarizing, $\delta_{r,t} = E(\Delta_{t,t}|\mathfrak{F}_r)$ is a number whereas $(E(\Delta_{t,t}|\mathfrak{F}_s))|\mathfrak{F}_r$ is a random variable. In the forecasting context, period s forecasts (e.g. $E(\Delta_{t,t}|\mathfrak{F}_s)$) are influenced by information gathered in periods $\{r+1, \dots, s\}$. Consequently, in period r , they are random variables.

Appendix C: Derivation of A1 and A3

We will show that assumptions (A1) and (A3) are consequences of forecasting by taking conditional expectations (i.e. $\delta_{s,t} = E(\Delta_{t,t}|\mathfrak{F}_s)$) and of normality of update vectors (I3).

Clearly, $E(\Delta_{t,t}|\mathfrak{F}_s) \in L^2(\Omega, \mathfrak{F}_s, P_s)$ where L^2 is the Hilbert space with the inner product $\langle f, g \rangle = E(fg)$. Let $\mathfrak{F}_s$ be the set of all functions measurable by time s . It is generated by updates observed before and in time s . Then, it is a subspace of L^2 . By the orthogonal projection theorem, $E(\Delta_{t,t}|\mathfrak{F}_s)$ has a unique representation

$$E(\Delta_{t,t}|\mathfrak{F}_s) = Pro\{E(\Delta_{t,t}|\mathfrak{F}_s) \text{ on } \mathfrak{F}_{s-1}\} + Pro\{E(\Delta_{t,t}|\mathfrak{F}_s) \text{ on } \mathfrak{F}_{s-1}^\perp\}$$

where Pro stands for projection and $\mathfrak{S}_{s-1}^\perp$ is the subspace orthogonal to $\mathfrak{S}_{s-1}$. Note that $Pro\{X \text{ on } \mathfrak{S}\} = E(X|\mathfrak{S})$, see definition 2.7.2. in [21]. Then, by theorem 34.4. in [25],

$$Pro\{E(\Delta_{t,t}|\mathfrak{S}_s) \text{ on } \mathfrak{S}_{s-1}\} = E(E(\Delta_{t,t}|\mathfrak{S}_s)|\mathfrak{S}_{s-1}) = E(\Delta_{t,t}|\mathfrak{S}_{s-1})$$

Note that $U_{s,t} \in \mathfrak{S}_{s-1}^\perp$ because

$$\{E(\Delta_{t,t}|\mathfrak{S}_s) - E(\Delta_{t,t}|\mathfrak{S}_{s-1})\} = Pro\{E(\Delta_{t,t}|\mathfrak{S}_s) \text{ on } \mathfrak{S}_{s-1}^\perp\}.$$

Since constant functions are measurable in any σ -field, i.e., $1 \in \mathfrak{S}_{s-1}$, and $U_{s,t} \in \mathfrak{S}_{s-1}^\perp$, $E(U_{s,t} \cdot 1) = 0$. This proves (A3). Combining (A3) with $E(U_{s,t}U_{s-1,t}) = 0$ yields (A1). With (I3), zero correlation would imply independence of $U_{s,t}$ and any random variable in $\mathfrak{S}_r$ for some $r < s$. Thus $U_{s,t}|\mathfrak{S}_r = U_{s,t}$ for $r < s$.

Appendix D: Forecasts with Lag Bias

Let us focus on a product p , tec and drop p , tec indices. Suppose that perceived age forecast are lag biased, i.e. as opposed to Equation 6

$$\delta_{s,t} = E(\Delta_{t,t}|\mathfrak{S}_s) + lb(t-s)$$

where $lb(t-s)$ denotes the deterministic bias as a function of the lag $t-s$. By equation (4), $u_{s,t} = \delta_{s,t} - \delta_{s-1,t}$. Now because of the lag bias,

$$E(U_{s,t}) = E(U_{s,t}|\mathfrak{S}_{s-1}) = lb(t-s) - lb(t-s+1) \neq 0.$$

Thus, we need to estimate the expected value of perceived age updates in addition to covariances. Note that shifting a random variable does not affect its variability, so $U_{s,t}$ and $U_{s,t} - lb(t-s) + lb(t-s+1)$ have the same variances, while the second term has zero expected value. Then, covariance matrix of updates can be estimated with the techniques used before, except that in covariance estimations multiplications of deviations from the mean is divided by sample size minus one (as opposed to sample size). Thus, we can accommodate lag bias by making moderate changes to SeDFAM. Above argument also applies to lag-biased family forecasts $d_{s,t}^p$.

Acknowledgement

This work was supported by a grant from Semiconductor Research Cooperation under the task “Modelling Random Processes”. Authors would like to thank to Joseph L. Schafer for allowing us to use his EM software, and representatives of several semiconductor manufacturers for the information and data they made available to us.

References

- [1] Çakanyıldırım, M. and R.O. Roundy. (1999). *Demand forecasting and capacity planning in the semiconductor industry*. Technical Paper no: 1229, SORIE, Cornell University, NY.
- [2] Meixell, M.J. and S.D. Wu. (1998). *Scenario analysis of demands in a technology market using leading indicators*. Working paper, Dept. of Industrial and Manufacturing Systems Engineering, Lehigh University, Bethlehem, Pennsylvania.
- [3] Hamilton, J. D. *Time Series Analysis*. Princeton University Press, New Jersey.
- [4] Mahajan, V. and Y. Wind (1988). *New product forecasting models*. International Journal of Forecasting No 4. 341–358.
- [5] Murray, G and Silver, E. (1966). *A Bayesian analysis of the style goods inventory problem*. Management Science Vol 12 No 11. 785–797.
- [6] Chang, S.H. and Fyffe, D.E. (1971) *Estimation of forecast errors for seasonal-style-goods sales* Management Science Vol 18 No 3. 89–96.

- [7] Bass, F. (1969). *A New product growth model for consumer durables*. Management Science Vol 15 No 5. 215–227.
- [8] Norton, J. and F. Bass (1987). *A Diffusion theory model of adoption and substitution for successive generations of high technology products*. Management Science Vol 33 No 9. 1069–1086.
- [9] Kurawarwala, A.A. and Matsuo, H. (1996). *Forecasting and inventory management of short life cycle products*. Operations Research Vol 44 No 1. 131–150.
- [10] Guerrero, V.M. and Elizondo, J.A. (1997). *Forecasting a cumulative variable using its partially accumulated data*. Management Science Vol 43 No 6. 879–889.
- [11] Kekre, S., Morton, T. and Smunt, T. (1990). *Forecasting using partially known demands*. International Journal of Forecasting No 6. 115–125.
- [12] Bodily, S.E. and J.R. Freeland (1988). *A simulation of techniques for forecasting shipments using firm orders-to-date*. Journal of Operational Research Society Vol 39 No 9. 833–846.
- [13] Azoury, K. (1985). *Bayes solution to dynamic inventory models under unknown demand distribution*. Management Science Vol 31 No 9. 1150–1160.
- [14] Bunn, D.W. and A.A. Salo. *Forecasting with scenarios*. European Journal of Operational Research No 68. 291–303.
- [15] Angelus, A., Porteus, E.L. and Wood, S.C. (1997) *Optimal sizing and timing of capacity expansions with implications for modular semiconductor wafer fabs*. Research Paper No.1479. Graduate School of Business, Stanford University.
- [16] Graves, S.C., H.C. Meal, S. Dasu and Y. Qui (1986). *Two stage production planning in a dynamic environment*. In Multi-Stage Production Planning and Inventory Control. S. Axsäter, C. Schneeweiss and E. Silver, (eds.), Lecture notes in economics and Mathematical systems, Springer-Verlag, Berlin, 266: 9–43.
- [17] Heath, D. and Jackson, P. (1994). *Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems*. IIE Transactions Vol 26 No 2. 17–30.
- [18] Güllü, R. (1996). *On the value of information in dynamic production inventory problems under forecast evolution*. Naval Research Logistics Vol 43. 289–303.
- [19] Toktay, L.B. (1998). *Analysis of a production-inventory system under a stationary demand process and forecast updates*. Unpublished Ph.D. dissertation, Operations Research Center, MIT, Cambridge, MA.
- [20] Graves, S.C., D.B. Kletter and W.B. Hetzel (1998). *A dynamic model for requirements planning with application to supply chain optimization*. Operations Research Vol.46 Supp. No.3: S35–S49.
- [21] Brockwell, P. and R.A. Davis. (1987). *Time Series: Theory and Methods*. Springer-Verlag New York Inc.
- [22] Anderson, T.W. (1984). *An introduction to multivariate statistical analysis*. John Wiley & Sons, Inc. New York.
- [23] Schafer, J.L. (1996). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.
- [24] Bickel, P.J. and K.A. Doksum (1977). *Mathematical Statistics*. Simon & Schuster Company, New Jersey.
- [25] Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, Inc. New York.