INSEAD
The Business School
for the World®

Working Paper Series

2016/48/DSC/TOM

Assessing Uncertainty from Point Forecasts

Anil Gaba
INSEAD, anil.gaba@insead.edu

Dana G. Popescu
INSEAD, dana.popescu@insead.edu

Zhi Chen
INSEAD, zhi.chen@insead.edu

The paper develops a model for combining point forecasts into a probability distribution for a variable of interest. Our approach allows for point forecasts to be correlated and admits uncertainty on the distribution parameters given the forecasts. Further, it provides an easy way to compute an augmentation factor needed to equate the dispersion of the point forecasts to that of the predictive distribution, which depends on the correlation between the point forecasts and on the number of forecasts. We show that ignoring dependence between point forecasts or parameter uncertainty can lead to assuming an unrealistically narrow probability distribution. We further illustrate the implications in a newsvendor context, where our model in comparison with other methods leads to an order quantity that has higher variance but is biased in the less costly direction, and generates an increase in expected profit that can exceed 20%.

Keywords: Correlated Experts; Point Forecasts; Demand Forecasting; Newsvendor Model

Assessing Uncertainty from Point Forecasts

Anil Gaba
Department of Decision Sciences, INSEAD

Dana G. Popescu
Department of Technology and Operations Management, INSEAD

Zhi Chen
Department of Decision Sciences, INSEAD

The paper develops a model for combining point forecasts into a probability distribution for a variable of interest. Our approach allows for point forecasts to be correlated and admits uncertainty on the distribution parameters given the forecasts. Further, it provides an easy way to compute an augmentation factor needed to equate the dispersion of the point forecasts to that of the predictive distribution, which depends on the correlation between the point forecasts and on the number of forecasts. We show that ignoring dependence between point forecasts or parameter uncertainty can lead to assuming an unrealistically narrow probability distribution. We further illustrate the implications in a newsvendor context, where our model in comparison with other methods leads to an order quantity that has higher variance but is biased in the less costly direction, and generates an increase in expected profit that can exceed 20%.

Key words: correlated experts, point forecasts, demand forecasting, newsvendor model

1. Introduction

Predicting an unrealized future variable is a continuous human endeavor, in part because it is crucial for shaping the decisions that we must make. An investor or a financial institution might be interested in predicting a future currency exchange rate for its hedging or trading strategy, or a retailer might be interested in predicting the demand for a new upcoming product to determine the order quantity. At times, a decision maker might have access to plentiful and relevant historical data such that robust statistical models could be established (e.g., electricity demand in a region). However, in many instances, even in the presence of much past data, an overlay of human judgment is inevitable due to an evolving context with rapidly changing conditions. In predicting the demand for a new fashion product, for example, one must account for not only issues such as rapidly changing tastes and competing products, but also the possibility of inducing demand for the product that might not otherwise exist. In the case of insufficient historical data, several new approaches involving, for example, artificial neural networks, fuzzy logic, machine learning, have been proposed. Despite these, subjective human judgment remains a key element in predictions across numerous real-life settings (Seifert et al. 2015).

Subjective forecasts for a variable often come in the form of point forecasts, assessments of complete probability distributions being a difficult cognitive task for even the most well-trained in probabilities and practically next to impossible or very noisy at the least in real-life settings with people not trained in probabilities. In one of the well-known industry practices in operations management, experts (such as sales people, designers, product managers, and the like) are used to provide point forecasts for the demand of upcoming new products. These forecasts then become a key input in inventory decisions such as order quantities for new products (Fisher and Raman 1996). This, however, involves an additional step by the decision maker, to convert the point forecasts for a variable into an estimate of uncertainty (such as a probability distribution) for that variable, which is the broad focus of this paper.

Consider a decision maker who might judge that the demand for a new product is likely to follow a normal distribution, but might have little or no information on the parameters of the distribution such as the mean and variance of the distribution. The decision maker then obtains point forecasts from k experts. It makes sense that the mean and dispersion of the k point forecasts are indicative of the mean and variance of the demand distribution. Fisher and Raman (1996), for example, use the mean of k forecasts as an estimate of the expected (mean) demand and augment the observed standard deviation of the k forecasts by a factor of 1.75 to create an estimate of the standard deviation of demand. The augmentation factor of 1.75 is justified by a calibration exercise with past data. In a similar setting in one of the widely taught cases in operations management, *Sport Obermeyer* (Hammond and Raman 1994), an augmentation factor of 2 is suggested. Gaur et al. (2007), using historical data on three different data sets, test the hypothesis that the variance of demand is a positively correlated with both the mean and the dispersion of the point forecasts. However, several questions remain. First, it is not conceptually clear as to how the augmentation factor arises and what it depends upon. Second, any past data used to estimate the augmentation factor might not be applicable for a new product, such as a new fashion item or short life-cycle product. Third, while the point forecasts provide relevant information on the distribution parameters, those are unlikely to completely eliminate uncertainty on the parameters. In this sense, the heuristics mentioned above are only *certainty-equivalent* approaches, as if the parameters of the demand distribution are certain once their point estimates are obtained from the point forecasts. This is of course not the case. These are some of the specific questions we attempt to answer in this paper.

Broadly speaking, we view our contribution as twofold. We first develop an approach that is easily tractable, and hence practical, for combining point forecasts into a probability distribution for a variable of interest. Our approach allows for the forecasts to be correlated and admits uncertainty about the distribution parameters given the forecasts. Second, we show that the augmentation factor mentioned above is contingent upon the dependence between the forecasts and on the number of forecasts. Although we refer to information sources for the point forecasts as human experts, our approach is equally valid for any type of information sources, such as models. We discuss all this in finer details below.

Despite all the efforts to create a group of independent experts, some form of stochastic dependence between their forecasts is inevitable. For example, Winkler (1981) notes pairwise correlations of sportwriters' errors of prediction in the range of 0.84 to 0.97, suggesting that experts might have similar training and experience, might have access to the same data, and might use similar aids (such as models) for their predictions. Similarly, Ashton (1986) observes an average correlation of 0.6 between business sales forecasts by managers. Fisher and Raman (2010) observe in a forecasting deliberation process that "one of the participants was more articulate and assertive than the others. Often, she swayed her colleagues, so the final decisions represented her preferences rather than the collective wisdom." It is intuitive that if the point forecasts are highly correlated, then their dispersion would tend to underestimate the uncertainty about the variable of interest, perhaps substantially so. A greater degree of dependence between experts would then entail a larger augmentation factor. Further, number of experts has a bearing on the parameter uncertainty. For example, 5 forecasts instead of 20 forecasts imply greater uncertainty on the parameters estimated. A Bayesian approach, in contrast to a certainty-equivalent approach, accounts for such parameter uncertainty.

Bayesian models for combining correlated point forecasts under normality have been developed, for example, by Geisser (1965), Winkler (1981), and Clemen and Winkler (1985), where the latter also illustrate loss of information due to dependence between the information sources. A necessary input in these models is a covariance matrix consisting of all pairwise correlations between the experts (measures of dependence between experts) and variances of each point forecasts (measures of accuracy of the experts). Often, there might be little or no past applicable data on the covariance matrix given a unique forecasting context. Further, there is frequently only one k-variate observation in the sample, such as k experts providing one forecast each for a novel product, which then contains no information content on either the pairwise correlations or the variances of the experts' forecasts. As a result, assessing the necessary prior parameters for the covariance matrix entails a daunting task in real life. For example, with k experts, $k + k(k-1)/2$ prior parameters need to be assessed on which the sample contains no information. Hence, practical applications of such models have been limited. Moreover, several papers have raised concerns around robustness of such models (Bunn 1985, Winkler and Clemen 1992, Chhibber and Apostolakis 1993). Clemen (1989), in a comprehensive review of the literature on combining forecasts, points out that such a model in real-life forecasting situations has had somewhat mixed results, along with a central finding that simple averaging of point forecasts often outperforms more complex methods and is an easy and a fairly robust way in terms of accuracy to predict an unrealized variable. This is reiterated in similar review on combining forecasts by Armstrong (2001). Schmittlein et al. (1990) explore the tradeoff in a model between added estimation error from using an additional parameter and the reduction in modeling misspecification associated with the additional parameter. Using simulations, they compare a base model with equal variances and zero correlations with alternative models that combine equal or unequal variances with a common zero or a non-zero correlation. The general spirit of their results is consistent with the intuition that one should use different variances for different experts only when the variances are expected to be "far" from the base case, with the threshold of "far" depending of course among other things on the amount of information available for estimation of the parameters. At the same time, they conclude that "if there is one consistent finding from empirical studies of multiple forecasts, it is that the forecasts are not uncorrelated."

Overall, there appears to be wide support in the literature for an equal-weights model (i.e., treating all experts equally) unless there is a strong reason to believe otherwise. In many real-life settings, an equal-weights model might often be a good approximation, especially where there is no way to confidently distinguish between the information content or capabilities of the different experts.

Building on these ideas, we develop a Bayesian approach with *exchangeable* experts (all experts are considered to have the same variance and a common correlation). This minimizes the assessment of prior parameters on which there is no information content in the sample to only one (the common correlation) as opposed to $k + k(k-1)/2$. We differ from earlier Bayesian aggregation models in two respects. One, instead of assuming the covariance matrix to be known or unknown, we assume the covariance matrix to be partially known (i.e., with unknown common variance but known common correlation). Two, we extend our approach also to the lognormal case, which to our knowledge has not been explored before.

We begin with a normal model, and provide a simple way to compute the augmentation factor needed to equate the standard deviation of the predictive distribution to the observed standard deviation of the point forecasts. We show that this augmentation factor depends on correlation between the point forecasts and on the number of experts, and could be as high as 6 or as low as 1. Given a number of experts, a higher correlation between point forecasts leads to a higher augmentation factor, implying greater uncertainty in the predictive distribution. This is consistent with Clemen and Winkler (1985), where they show that there is a loss of information due to dependence between experts. On the other hand, given a correlation between experts, a higher number of experts leads to lower uncertainty in the predictive distribution, resulting from lower parameter uncertainty, but only up to a limit. In other words, if there is dependence between experts, the resulting loss of information cannot be compensated for by simply increasing the number of experts. We compare our approach with other methods used, for example, in the operations management literature, and highlight the impact of ignoring any dependence between the experts or the parameter uncertainty or both. We extend this analysis to a decision making context of a newsvendor setting, and show the impact of our approach on the order quantity and expected profit. Our model in comparison with the other methods leads to an order quantity that is biased in the less costly direction with a higher variance. At the same time, our model leads to an increase in expected profit that can exceed 20%. We investigate our model with uncertainty on the common correlation between experts and with heterogeneity in the pairwise correlations between experts, and show that the augmentation factor in our model is fairly robust in these respects. Finally, we extend our approach under normality to the case where the variable of interest and the point forecasts have a lognormal distribution, and show that the results under the normal model only get exacerbated.

Our approach is easy to implement in practice and adds to a broader and growing stream of research on forecasting in the operations management field. For example, to name a few, Kremer et al. (2015) consider aggregate vs. sum of bottom-up forecasts for a firm and show that the correlation structure between the bottom-up forecasts has informational value. Shumsky (1998) explores optimal updating for forecasts for future events. Özer et al. (2011) investigate credible forecast sharing between a supplier and a manufacturer. And, Schweitzer and Cachon (2000) illustrate some behavioral biases in order quantity decisions.

The rest of the paper is organized as follows. In §2, we develop our model under normality. In §3, we illustrate our results in a newsvendor settings. In §4, we investigate the robustness of our model with respect to uncertainty and heterogeneity in the dependence between experts, and extend our approach to the lognormal case. §5 follows with a summary and discussion. All proofs are provided in the Appendix.

2. A Model for a Probability Distribution from Point Forecasts

Let a random variable of interest to a decision maker be $\tilde{y}$, which for example could be a future observation of demand for a new product. Suppose the decision maker models the probability distribution of $\tilde{y}$ conditional on a parameter (or a vector of parameters) θ with density function $f(\tilde{y}|\theta)$, with θ unknown. For example, $f(\tilde{y}|\theta)$ might be modeled as a normal density function conditional on $\theta = (\mu, \sigma^2)$, where μ and σ^2 are unknown mean and variance respectively of the normal distribution. Further, let a probability density function $h(\theta)$ reflect the decision maker's prior uncertainty on θ. Now suppose that the decision maker obtains some additional information before realization of $\tilde{y}$. We consider such information in the form of k point forecasts $\boldsymbol{x} = (x_1, x_2, ..., x_k)'$ of $\tilde{y}$ from k different sources. Then, given $\boldsymbol{x}$, the predictive distribution for $\tilde{y}$ is given by

$$f(\tilde{y}|\boldsymbol{x}) = \int_{\Theta} f(\tilde{y}|\theta)h(\theta|\boldsymbol{x})d\theta, \tag{1}$$

where

$$h(\theta|\boldsymbol{x}) \propto h(\theta)l(\boldsymbol{x}|\theta) \tag{2}$$

is the posterior distribution of θ given $\boldsymbol{x}$ and $l(\boldsymbol{x}|\theta)$ is the likelihood function for $\boldsymbol{x}$ given θ. In this setup, $\tilde{y}$ and $\boldsymbol{x}$ are conditionally independent given θ, and any dependence between the point estimates $x_1, ..., x_k$ is included in $l(\boldsymbol{x}|\theta)$. If the decision maker does not have any prior information on θ, then a diffuse (flat) prior on θ can be used. On the other hand, if the decision maker does possess some prior information (such as relevant experience or past data), then that information should be included in $h(\theta)$. Alternatively, decision maker's prior information can also be modeled as there being $k+1$ experts instead of k experts and then assuming $h(\theta)$ to be diffuse. Whether the decision maker's information is included in $h(\theta)$ or in the form of $(k+1)^{th}$ expert hinges upon how to best model the dependence structure within experts and between the experts and the decision maker. The predictive distribution for $\tilde{y}$ in (1) accounts for two types of uncertainty, the uncertainty of $\tilde{y}$ given θ and the uncertainty about the parameter θ given $\boldsymbol{x}$. This is a typical Bayesian approach for aggregating expert opinions.

Dependence between experts is often modeled either as correlation between the experts' predictions (Lichtendahl Jr et al. 2013) or as correlation between the errors of experts predictions (Winkler 1981). We take the former approach, i.e., we define dependence between two experts in terms of the correlation between their predictions, i.e., $Corr[x_i, x_j|\theta] = \rho_{ij} = \rho$, $i \neq j$. However, our model can be easily adapted for other definitions of dependence, such as for correlations between the errors of experts' forecasts.

We extend this approach below under normality.

2.1. A Normal Model

Suppose that $\tilde{y} \sim N(\mu, \sigma^2)$, with μ and σ^2 unknown. The k experts provide point forecasts $\boldsymbol{x} = (x_1, ..., x_k)'$ for $\tilde{y}$ that follow a multivariate normal with a mean vector $\boldsymbol{\mu} = \mu\boldsymbol{e}$, where $\boldsymbol{e} = (1, ..., 1)'$ is a $k \times 1$ column vector, and a $k \times k$ positive definite covariance matrix $\boldsymbol{\Sigma}$ with diagonal elements σ^2 and off-diagonal elements $\rho\sigma^2$. This implies that the experts are unbiased and exchangeable, and that each expert receives a signal from the demand distribution.[1] Additionally, we assume that ρ is known. Later, in §4, we introduce uncertainty and heterogeneity with respect to ρ.

[1] Any deviation from this assumption does not the change the overall nature of the results in terms of impact of correlation between experts.

The likelihood function for the k-variate forecast $\boldsymbol{x}$ is given by

$$l(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{3}$$

Setting $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Sigma}_\rho$, where $\boldsymbol{\Sigma}_\rho$ is a $k \times k$ is a matrix with diagonal elements 1 and off-diagonal elements ρ, and precision $\lambda = 1/\sigma^2$, the likelihood function can be rewritten as

$$l(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = l(\boldsymbol{x}|\boldsymbol{\mu},\lambda) \propto \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_\rho^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{4}$$

Lemma 1 below further simplifies the likelihood function in (4).

Lemma 1. *The likelihood function in* (4) *can be represented as*

$$l(\boldsymbol{x}|\mu,\lambda) \propto \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda k^*}{2}(\mu-\overline{x})^2\right) exp\left(-\frac{\lambda}{2}(k-1)s^{*2}\right), \tag{5}$$

where $\overline{x} = (1/k)\sum_{i=1}^k x_i$ and $s^2 = \sum_{i=1}^k (x_i-\overline{x})^2/(k-1)$ are respectively the sample mean and the sample variance obtained from the k point forecasts, $k^* = k/(1+(k-1)\rho)$, and $s^{*2} = s^2/(1-\rho)$ is an unbiased estimator of σ^2.

We model the decision maker's prior information on μ and λ with a normal-gamma (NG) distribution which is a natural conjugate prior distribution for a normal process with unknown mean and variance (for our parameterization, see Hoff 2009):

$$\begin{aligned} f(\mu,\lambda) &= NG(\mu,\lambda|\mu_0,n_\mu,v_0,n_v) \\ &= f(\mu|\lambda)f(\lambda) = N(\mu|\mu_0,(n_\mu\lambda)^{-1})Ga(\lambda|\frac{n_v}{2},\frac{n_v v_0}{2}) \\ &\propto \lambda^{\frac{1}{2}} exp\left(-\frac{n_\mu\lambda(\mu-\mu_0)^2}{2}\right)\lambda^{\frac{n_v}{2}-1} exp\left(-\frac{n_v v_0}{2}\lambda\right). \end{aligned} \tag{6}$$

The NG prior in (6) can reflect a wide variety of information regarding μ and λ. *A priori*, the conditional distribution of μ given λ is normal with mean μ_0 and variance $(n_\mu\lambda)^{-1}$, where $\mu_0 \in \mathbb{R}$ and $n_\mu > 0$. And, the marginal distribution of λ is gamma with shape parameter $n_v/2$ and rate parameter $n_v v_0/2$ for some $n_v > 0$ and $v_0 > 0$, so that $E(\lambda) = 1/v_0$ and $Var(\lambda) = 2/(n_v v_0^2)$. With this parametrization, one can say that *a priori* the decision maker's best guess of μ is μ_0 and n_μ can be viewed as the equivalent sample size for the prior information on μ. And, the decision maker's best guess of $\lambda = 1/\sigma^2$ is $1/v_0$ with n_v as the equivalent sample size for the prior information on λ. Note that with the prior on λ in (6), σ^2 has an inverse-gamma distribution with shape parameter $n_v/2$ and rate parameter $n_v v_0/2$, so that $E(\sigma^2) = (n_v/(n_v-2))v_0$ and $Var(\sigma^2)$ is decreasing in n_v.

Theorem 1. *Consider $\tilde{y} \sim N(\mu,\sigma^2)$, with μ and σ^2 unknown. Suppose that k experts provide point forecasts $\boldsymbol{x} = (x_1,...,x_k)'$ for $\tilde{y}$ that follow a multivariate normal with a mean vector $\boldsymbol{\mu} = \mu\boldsymbol{e}$, where $\boldsymbol{e} =$*

$(1,...,1)'$ *is a* $k \times 1$ *column vector, and a* $k \times k$ *positive definite covariance matrix* $\mathbf{\Sigma}$ *with diagonal elements* σ^2 *and off-diagonal elements* $\rho\sigma^2$*, with the common correlation* ρ *known. Then, with the normal-gamma prior on* μ *and* σ *in (6) and given a sample result of* $\bar{x}$ *and* s^2 *defined in Lemma 1,*

a) $f(\mu,\lambda|\boldsymbol{x})$ *is also a normal-gamma distribution of the same form as in (6) but with updated parameters:*

$$\begin{aligned} f(\mu,\lambda|\boldsymbol{x}) &= NG(\mu,\lambda|\mu^*,n_\mu^*,v^*,n_v^*) \\ &= f(\mu|\lambda,\boldsymbol{x})f(\lambda|\boldsymbol{x}) = N(\mu|\mu^*,(n_\mu^*\lambda)^{-1})Ga(\lambda|\frac{n_v^*}{2},\frac{n_v^* v^*}{2}), \end{aligned} \tag{7}$$

where $n_\mu^* = n_\mu + k^*$, $\mu^* = \frac{n_\mu \mu_0 + k^* \bar{x}}{n_\mu^*}$, $n_v^* = n_v + k$ *and* $v^* = \frac{1}{n_v^*}\left(n_v v_0 + (k-1)s^{*2} + \frac{n_\mu k^*}{n_\mu + k^*}(\bar{x} - \mu_0)^2\right)$*, with* k^* *and* s^{*2} *as defined in Lemma 1.*

b) For $\rho = 0$, $f(\mu,\lambda|\boldsymbol{x})$ *is the same as with k independent experts.*

c) For $\rho > 0$, $f(\mu|\lambda,\boldsymbol{x})$ *is the same as with* k^* *independent experts, where* $k^* \leq k$ *is a decreasing function of* ρ*. And,* $f(\mu,\lambda|\boldsymbol{x})$ *is the same as if a sample variance of* s^{*2} *rather than* s^2 *is observed with k independent experts, where* $s^{*2} \geq s^2$ *is an increasing function of* ρ*, while at the same time as if a sample mean of* $\bar{x}$ *is observed with only* k^* *of k independent experts.*

Theorem 1, Part (a), states that posterior distribution of μ and λ is of the same natural-conjugate form as the prior. Part (b) of the theorem shows that for $\rho = 0$ the posterior distribution of μ and λ is the same as one would obtain with k independent observations. Finally, Part (c) shows that for $\rho > 0$ the posterior conditional distribution of μ given λ is the same as one would obtain with k^* independent experts. In that sense, k^* can be viewed as an equivalent independent sample size for inferences about μ given λ. Note that $k^* = k/(1+(k-1)\rho)$. It is easy to see that $k^* = k$ for $\rho = 0$. And, as ρ gets larger than zero, k^* shrinks, decreasing to 1 for $\rho = 1$. This is consistent with the results in Clemen and Winkler (1985) for the case of known λ, that any positive dependence between the point forecasts reduces the information content of the forecasts for inferences about μ. Further, while there is loss on information on μ, there is no loss of information on λ. The sample variance s^2 is simply adjusted by factor $1/(1-\rho)$ to obtain an unbiased estimator s^{*2} of σ^2.

It follows from Theorem 1 that the marginal posterior distribution of μ, $f(\mu|\boldsymbol{x})$, is a t distribution with n_v^* degrees of freedom, location μ^* and scale $\sqrt{v^*/n_\mu^*}$, such that $E(\mu|\boldsymbol{x}) = \mu^*$ for $n_v^* > 1$ and $Var(\mu|\boldsymbol{x}) = (n_v^*/(n_v^*-2))v^*/n_\mu^*$, for $n_v^* > 2$.

The posterior mean of μ is a weighted mixture of the prior mean μ_0 and the sample mean $\bar{x}$, where the weights depend on the relative strengths of the prior information on μ (i.e., n_μ) and the sample information on μ (i.e., k^*). A higher ρ decreases the weight on the sample information. Further, $E(\lambda|\boldsymbol{x}) = 1/v^*$, where v^* is a weighted mixture of v_0 and s^{*2} (the sample variance adjusted for known ρ), and an additional term that takes into account the difference between the prior mean of μ (i.e., μ_0) and the sample mean (i.e., $\bar{x}$). A higher ρ shifts the posterior distribution towards lower values of λ, but the coefficient of variation remains the same regardless of ρ. In that sense, there is no loss of information with respect to λ.

The decision maker's primary interest is in the predictive distribution for $\tilde{y}$ given $\boldsymbol{x}$, which is shown in Corollary 1.

COROLLARY 1. *The predictive distribution for $\tilde{y}|\boldsymbol{x}$ is a t distribution with degrees of freedom n_v^*, location parameter μ^*, and scale parameter $\sqrt{(n_\mu^*+1)v^*/n_\mu^*}$, so that*

$$E(\tilde{y}|\boldsymbol{x}) = \mu^*, \textit{for } n_v^* > 1, \textit{and} \tag{8}$$

$$Var(\tilde{y}|\boldsymbol{x}) = Var(\tilde{y}|\mu,\boldsymbol{x}) + Var(\mu|\boldsymbol{x}) = \frac{n_v^*}{n_v^*-2}\left(v^* + \frac{v^*}{n_\mu^*}\right), \textit{for } n_v^* > 2. \tag{9}$$

The first term in (9), $Var(\tilde{y}|\mu,\boldsymbol{x})$, corresponds to the sampling uncertainty given μ. And, the second term, $Var(\mu|\boldsymbol{x})$, reflects the uncertainty about μ itself. The uncertainty about the precision λ is embedded in $Var(\tilde{y}|\mu,\boldsymbol{x})$ and $Var(\mu|\boldsymbol{x})$. All else unchanged, a higher ρ leads to a greater loss of information about μ in the sample (i.e., the forecast) and a lower expectation of λ given the sample (i.e., increases v^*), resulting in a higher variance for $\tilde{y}|\boldsymbol{x}$. On the other hand, all else held equal with $\rho > 0$, a higher k reduces the parameter uncertainty, although much less for μ than for λ, resulting in a lower variance for $\tilde{y}|\boldsymbol{x}$.

Often, the decision maker might have very little or no any prior information on μ and λ, and must rely entirely upon the information provided by the k experts. In that case, the predictive distribution is much simplified, which we explore in much greater detail below.

2.1.1. Diffuse Prior Information on μ and σ^2 . In this section, we focus on the case of the predictive distribution for $\tilde{y}|\boldsymbol{x}$ with a diffuse prior on μ and λ and compare our model with some other purely data-based methods. More specifically, we compare four methods described below.

Predictive Distribution with Known ρ (*PD*). This is our model, with ρ known, and a diffuse prior for μ and λ. A common choice for a diffuse prior is a NG prior in (6) with parameters $n_\mu = n_v = 0$, which is an improper prior and also the Jeffreys prior. With this diffuse prior, $f(\mu,\lambda|\boldsymbol{x}) = NG(\mu,\lambda|\mu^*,n_\mu^*,v^*,n_v^*)$ with $n_\mu^* = k^*, \mu^* = \bar{x}, n_v^* = k$, and $v^* = (k-1)s^{*2}/k$. Then, $f(\tilde{y}|\boldsymbol{x})$ is a t distribution with k degrees of freedom, location $\bar{x}$, and scale $\sqrt{(k^*+1)v^*/k^*}$, which yields $E(\tilde{y}|\boldsymbol{x}) = \overline{x}$ and

$$Var(\tilde{y}|\boldsymbol{x}) = \frac{k-1}{k-2}\left(\frac{1+\rho}{1-\rho} + \frac{1}{k}\right)s^2. \tag{10}$$

The $Var(\tilde{y}|\boldsymbol{x})$ in (10) is larger for a higher ρ given k, reflecting a greater loss of information due to higher correlation between the experts. On the other hand, it is smaller for a higher k given ρ, reflecting reduced parameter uncertainty. For very large k, $Var(\tilde{y}|\boldsymbol{x}) \approx (1+\rho)s^2/(1-\rho)$.

Further, with also $\rho = 0$, $Var(\tilde{y}|\boldsymbol{x}) \approx s^2$. In other words, with k very large and $\rho = 0$, $f(\tilde{y}|\boldsymbol{x})$ converges to a normal distribution with mean $\bar{x}$, and variance s^2. With respect to the hypothesis formulated in Gaur et al. (2007), the variance of the predictive is indeed positively correlated with the dispersion of the point forecasts, however it does not depend on the mean of the point forecasts.

Predictive Distribution with $\rho \overset{\text{set}}{=} 0$ (PD_0). This is the same as PD above but with ρ assumed to be 0. In other words, this approach retains parameter uncertainty, but assumes independence between the experts. In this setup, $E(\tilde{y}|\boldsymbol{x}) = \overline{x}$ and

$$Var(\tilde{y}|\boldsymbol{x}) = \frac{k-1}{k-2}\left(1 + \frac{1}{k}\right) s^2. \tag{11}$$

While ignoring ρ does not impact $E(\tilde{y}|\boldsymbol{x})$, $Var_{PD_0}(\tilde{y}|\boldsymbol{x}) < Var_{PD}(\tilde{y}|\boldsymbol{x})$ for any $\rho > 0$, with $Var_{PD}(\tilde{y}|\boldsymbol{x})$ being larger by a factor of $((1+\rho)/(1-\rho) + 1/k)/(1+1/k)$. Even in the limit with a very large k, $Var_{PD}(\tilde{y}|\boldsymbol{x})$ remains larger by a factor of $(1+\rho)/(1-\rho)$. The dependence among experts (i.e., $\rho > 0$) causes a loss of information about $\tilde{y}$, and even a very large k (number of experts) can not overcome this. In other words, ignoring ρ leads to spurious accuracy in the predictions.

Certainty Equivalent Method with ρ Known (CE). Here, it is assumed that $\tilde{y}|\boldsymbol{x} \sim N(\overline{x}, s^{*2} = s^2/(1-\rho))$. In this approach, $\overline{x}$ is used as a point estimate of μ, and s^{*2} (which is an unbiased estimator of σ^2) is used as a point estimate for σ^2. While this approach incorporates the known correlation between experts, it ignores any parameter uncertainty once the point estimates of the parameters are obtained. In this sense, this is a certainty equivalent model corresponding to PD. $E(\tilde{y}|\boldsymbol{x})$ is of course not impacted, but as one would expect $Var_{CE}(\tilde{y}|\boldsymbol{x}) < Var_{PD}(\tilde{y}|\boldsymbol{x})$ for any $k > 2$, with

$$\frac{Var_{PD}(\tilde{y}|\boldsymbol{x})}{Var_{CE}(\tilde{y}|\boldsymbol{x})} = \frac{k-1}{k-2}\left[(1+\rho) + \frac{(1-\rho)}{k}\right]. \tag{12}$$

Even with $\rho = 0$, CE yields a lower variance for $\tilde{y}|\boldsymbol{x}$ than PD, as it ignores parameter uncertainty and assumes a normal instead of a t distribution.

Certainty Equivalent Method with $\rho \overset{\text{set}}{=} 0$ (CE_0). In this approach, $\tilde{y}|\boldsymbol{x} \sim N(\overline{x}, s^2)$. As in CE, this approach does not include any parameter uncertainty, and further ignores ρ. Note that $Var_{CE_0}(\tilde{y}|\boldsymbol{x})$ is the smallest of all the models discussed here, and in that sense includes the most spurious accuracy.

For ease of reading, Table 1 below summarizes our model along with the three other methods discussed above, and provides for each the distribution of $\tilde{y}|\boldsymbol{x}$ and the corresponding $Var(\tilde{y}|\boldsymbol{x})$, where $t_k(a,b)$ denotes a t distribution with degrees of freedom k, location a, and scale b. As mentioned above, $E(\tilde{y}|\boldsymbol{x}) = \overline{x}$ in all the four methods, but the variances differ depending on whether one accounts for the parameter uncertainty and the dependence between the experts.

Table 1 A summary of the four methods (PD, PD_0, CE, CE_0) in the normal model

Method	$f(\tilde{y}\|\boldsymbol{x})$	$Var(\tilde{y}\|\boldsymbol{x})$	
PD	$t_k(\overline{x}, \sqrt{\frac{k-1}{k}(\frac{1+\rho}{1-\rho}+\frac{1}{k})}s)$	$\frac{k-1}{k-2}\left(\frac{1+\rho}{1-\rho}+\frac{1}{k}\right)s^2$	includes ρ and parameter uncertainty
PD_0	$t_k(\overline{x}, \sqrt{\frac{k-1}{k}(1+\frac{1}{k})}s)$	$\frac{k-1}{k-2}\left(1+\frac{1}{k}\right)s^2$	ignores ρ but includes parameter uncertainty
CE	$N(\overline{x}, \frac{s^2}{1-\rho})$	$\frac{s^2}{1-\rho}$	includes ρ but ignores parameter uncertainty
CE_0	$N(\overline{x}, s^2)$	s^2	ignores ρ and parameter uncertainty

Figure 1 shows the standard deviation of $\tilde{y}|\boldsymbol{x}$ under the four methods as a function of ρ for $k =$ 3, 7 and 100, rescaled by setting the standard deviation under CE_0 for a given k equal to 1. Let SD_i be the standard deviation of $\tilde{y}|\boldsymbol{x}$ under method i. As discussed above, for any k and ρ, SD_{CE_0} is the smallest and SD_{PD} is the largest. Further, SD_{CE_0} and SD_{PD_0} are flat since these correspond to approaches that assume $\rho = 0$. Note that while both PD and CE incorporate dependence between experts, CE ignores uncertainty about the parameters. Hence, $SD_{CE} < SD_{PD}$. As k gets larger, SD_{CE} and SD_{PD} get closer, since parameter uncertainty is reduced. However, this happens only up to a limit (shown in (12)) due to the loss of information arising from the dependence between experts. For example, with $\rho = 0.6$, SD_{PD} as compared to SD_{CE} is about 86% larger for $k = 3$, 41% larger for $k = 7$, and still about 27% larger for $k = 100$. Comparing SD_{PD} and SD_{PD_0} reflects the effect of ignoring ρ while incorporating uncertainty about the parameters. The gap between the two becomes larger with a higher ρ. And, even with a very large k, SD_{PD} is larger by a factor of $(1+\rho)/(1-\rho)$ (shown in (10)). The two approaches PD_0 and CE_0 ignore ρ, but PD_0 incorporates parameter uncertainty. As k gets larger, there is reduction in parameter uncertainty, and for large enough k (for example, $k = 100$), SD_{PD_0} and SD_{CE_0} are almost identical. Overall, it is clear that ignoring ρ or the uncertainty about μ and σ^2 leads to underestimation of uncertainty about $\tilde{y}$. In many real-life settings, it might not be unusual for ρ to be moderately large (say, between 0.6 to 0.8), and k in the range of 5 to 10 is more the norm rather than the exception, in which case not taking into account ρ or the parameter uncertainty leads to serious underestimation of uncertainty about $\tilde{y}$. For example, with $k = 7$ and $\rho = 0.6$, the underestimation of SD is about 46% due to ignoring ρ, about 29% due to ignoring the parameter uncertainty, and about 55% due to ignoring both. The same percentages rise to 65%, 32%, and 70%, respectively, with $\rho = 0.8$.

Figure 1 **Standard deviation (SD) of $\tilde{y}|x$ under the four methods (PD, PD_0, CE, CE_0) in the normal model, rescaled with $SD_{CE_0} = 1$, as a function of ρ for selected values of k**

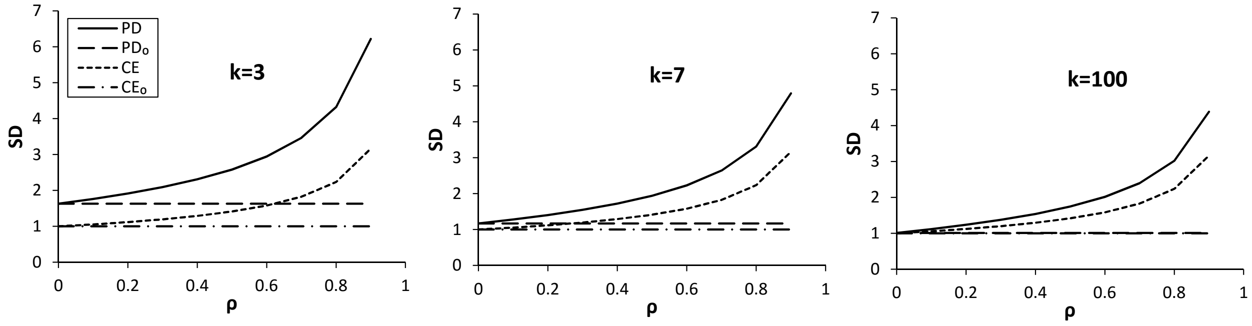


Table 2 below provides similar information. It shows an augmentation factor δ that is needed to equate the standard deviation of the k estimates (i.e, $s = \sqrt{\sum_{i=1}^{k}(x_i - \overline{x})^2/(k-1)}$) with the standard deviation of $\tilde{y}|x$ in our model (i.e., SD_{PD}) for different values of ρ and k. A popular heuristic in the operations management literature for assessing the uncertainty about $\tilde{y}|x$ has been a certainty equivalent model that assumes $\tilde{y}|x \sim N(\overline{x}, (\delta s)^2)$ with a choice of δ between 1.75 and 2 (Fisher and Raman 1996, Hammond and Raman 1994). The choice of δ is not motivated by the potential dependence between experts or the number of experts, but more in terms of calibration with past forecast errors. Further, while it has been acknowledged that the distribution of $\tilde{y}|x$ might be a t rather than a normal, this is motivated not by uncertainty about the parameters (μ and σ^2) but more by empirical data on past forecast errors. Our model provides a rationale for δ, with attribution to the dependence between experts and to the number of experts. In Table 2, δ decreases with a higher k given ρ, reflecting more information and hence lower uncertainty about the parameters. However, for $\rho > 0$, even with large k, δ is higher than 1, significantly so for high values of ρ. This is consistent with the earlier discussion about loss of information due to dependence between the experts, and a large k does not compensate for that. In fact, as shown in (10), in the limit, $\delta = (1+\rho)/(1-\rho)$. On the other hand, δ increases with a higher ρ given k, indicating greater loss of information about $\tilde{y}$ with greater dependence between the experts. For example, given $k = 6$, δ increases from 1.99 for $\rho = 0.5$ to 3.39 for $\rho = 0.8$. A value of $\delta = 1.75$ used in Fisher and Raman (1996) roughly corresponds to (k,ρ) pairs of, for example, (3, 0.1), (4, 0.3), (7, 0.4), and (100, 0.5). And, similarly, the $\delta = 2$ used in Hammond and Raman (1994) corresponds to, for example, (4, 0.4), (6, 0.5), and (100, 0.6).

3. An Illustration for the Newsvendor Problem

We illustrate some of the implications of our model in §2 in the decision making context of a newsvendor problem. In a typical newsvendor setting, a decision maker must make a one-time ordering decision ahead of the selling season without knowing the demand. The parameters for

Table 2 Augmentation factor δ in the normal model for selected values of k and ρ

	ρ									
k	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	1.63	1.76	1.91	2.09	2.31	2.58	2.94	3.46	4.32	6.22
4	1.37	1.49	1.62	1.78	1.97	2.21	2.52	2.98	3.72	5.37
5	1.26	1.38	1.51	1.66	1.84	2.07	2.37	2.80	3.50	5.06
6	1.21	1.32	1.44	1.59	1.77	1.99	2.28	2.70	3.39	4.89
7	1.17	1.28	1.40	1.55	1.72	1.94	2.23	2.64	3.31	4.79
8	1.15	1.25	1.38	1.52	1.69	1.91	2.19	2.60	3.26	4.72
9	1.13	1.23	1.36	1.50	1.67	1.89	2.17	2.57	3.23	4.67
10	1.11	1.22	1.34	1.48	1.65	1.87	2.15	2.55	3.20	4.64
20	1.05	1.16	1.28	1.42	1.59	1.79	2.07	2.46	3.09	4.48
100	1.01	1.12	1.24	1.37	1.54	1.74	2.01	2.39	3.02	4.38

the newsvendor problem are as follows: $c > 0$ is the unit cost, $p > c$ is unit selling price, and $v < c$ is the unit salvage value. Let $\Pi(y,q)$ be the newsvendor profit function, where y is the realized demand and q is the order quantity. For a given demand distribution, the solution to the newsvendor problem is $q^* = \arg\max_q E[\Pi(\tilde{y}|q)] = F_{\tilde{y}}^{-1}(CR)$, where $F_{\tilde{y}}$ is the cumulative distribution function of demand and $CR = (p-c)/(p-v)$ is the fractile of the demand distribution corresponding to the optimal ordering quantity (a.k.a. *critical ratio*).

We focus on a comparison of the four methods described in §2.1.1 for the normal model. We are interested primarily in order quantity and the profit associated with each of the demand estimation methods, and how these might compare to the same under perfect information about the demand distribution parameters. As before, k denotes the number of experts, $\bar{x}$ and s^2 denote the sample mean and the sample variance respectively of the k point forecasts. In our normal model, demand $\tilde{y} \sim N(\mu,\sigma^2)$. Table 1 in §2.1.1 summarizes the estimated demand distribution resulting from each of the four methods.

3.1. Impact of the Demand Estimation Method on the Newsvendor Order Quantity

Let $t_{CR,k} = T_k^{-1}(CR)$ and $z_{CR} = \Phi^{-1}(CR)$, where $T_k(\cdot)$ and $\Phi(\cdot)$ are respectively CDFs of the standard t distribution with k degrees of freedom and the standard normal distribution. Then, conditional on $\boldsymbol{x} = (x_1,...,x_k)'$, the optimal order quantities under each of the four methods is given in Table 3.

Table 3 Optimal order quantities under the four methods (PD, PD_0, CE, CE_0) given x in the normal model

$q_{PD} = \overline{x} + t_{CR,k} s \sqrt{\frac{k-1}{k}\left(\frac{1+\rho}{1-\rho} + \frac{1}{k}\right)}$
$q_{PD_0} = \overline{x} + t_{CR,k} s \sqrt{\frac{k-1}{k}\left(1 + \frac{1}{k}\right)}$
$q_{CE} = \overline{x} + z_{CR} s \sqrt{\frac{1}{1-\rho}}$
$q_{CE_0} = \overline{x} + z_{CR} s$

An immediate observation is that for all $\rho \geq 0$, we have $q_{PD} \geq q_{PD_0}$ and $q_{CE} \geq q_{CE_0}$ whenever $t_{CR,k}$, $z_{CR} \geq 0$ (i.e., $CR \geq 0.5$). And, we have $q_{PD} \leq q_{PD_0}$ and $q_{CE} \leq q_{CE_0}$ whenever $t_{CR,k}$, $z_{CR} \leq 0$ (i.e., $CR \leq 0.5$). Next, we compare q_{PD} to q_{CE}. Let $x_k(u)$ be the solution of the equation $T_k(x) = \Phi(u)$.

A simple lower bound on $x_k(u)/u$ can be obtained from the Cornish-Fisher expansion of $x_k(u)$ (see, e.g., Fujikoshi and Mukaihata 1993):

$$x_k(u) = u + \frac{1}{4k}(u^3 + u) + \frac{1}{96k^2}(5u^5 + 16u^3 + 3u) + ...$$

For all $u > 0$ and $k > 0$, we have $\frac{x_k(u)}{u} \geq 1 + \frac{1}{4k}$. Similarly, for $u < 0$ and $k > 0$, $\frac{x_k(u)}{u} \leq 1 + \frac{1}{4k}$. But, for $k > 2$ we have:

$$\left(1 + \frac{1}{4k}\right)\sqrt{\frac{k-1}{k}(\frac{1+\rho}{1-\rho} + \frac{1}{k})} \geq \sqrt{\frac{1}{1-\rho}},$$

which implies that $q_{PD} \geq q_{CE}$ for $CR \geq 0.5$ and $q_{PD} \leq q_{CE}$ for $CR \leq 0.5$.

The above results are conditional on $\boldsymbol{x}$, the k point forecasts. Let $\tilde{q}_i$ be the unrealized value of q_i for $i \in \{PD, PD_0, CE, CE_0\}$, i.e., before $\boldsymbol{x}$ is observed. Proposition 1 below provides some properties of the unconditional distributions of the order quantities under the four demand estimation methods.

Proposition 1. *The following inequalities hold:*

$$\begin{aligned}
E[\tilde{q}_{PD}] &\leq \min\{E[\tilde{q}_{PD_0}], E[\tilde{q}_{CE}], E[\tilde{q}_{CE_0}]\}, \textit{ for } CR \leq 0.5,\\
E[\tilde{q}_{PD}] &\geq \max\{E[\tilde{q}_{PD_0}], E[\tilde{q}_{CE}], E[\tilde{q}_{CE_0}]\}, \textit{ for } CR \geq 0.5,\\
Var[\tilde{q}_{PD}] &\geq \max\{Var[\tilde{q}_{PD_0}], Var[\tilde{q}_{CE}], Var[\tilde{q}_{CE_0}]\}, \textit{ for any } CR \in [0,1].
\end{aligned}$$

The order quantity under the PD method is the most conservative of the four, by having the greatest bias in the direction of the less costly consequence: underestimation of demand for $CR \leq 0.5$ and overestimation of demand for $CR \geq 0.5$. However, the variance of the order quantity is always highest under the PD model.

We also compare the order quantities under the four methods to the optimal order quantity q^* under perfect information about the demand distribution parameters (i.e., when μ and σ^2 are known). For such a comparison, we use simulations. For a given CR, CV, k, and ρ, we generate 100,000 k-variate observations of $\boldsymbol{x}$ from a multivariate normal and for each observation compute the order quantities under each of the four methods along with q^* (which does not depend on $\boldsymbol{x}$ and hence is the same across all observations). We do this for the cases of low (0.2) and high (0.8) CR, $CV = 0.2$ (assuming WLOG $\mu = 10$ and $\sigma = 2$), $k \in \{3, 7, 100\}$, and $\rho \in \{0, 0.1, ..., 0.9\}$. Figure 2 shows the expected order quantities, rescaled with $q^* = 1$, under the four methods for various levels of the parameters. Figure 3 shows, as an example, the complete distributions of the order quantity, rescaled after setting $q^* = 1$, under the four methods for the case of $CV = 0.2$, $k = 7$, and $\rho = 0.6$ with low CR (0.2) and high CR (0.8). Given k, a higher ρ means greater uncertainty about demand, and PD compensates for this by creating a greater bias in the order quantity relative to q^* in the direction of the less costly consequence (i.e., less than q^* for $CR \leq 0.5$ and greater than q^* for $CR \geq 0.5$). On the other hand, PD_0 and CE_0 lead to a bias relative to q^* in the opposite direction, the more costly direction. This arises due to underestimation of demand uncertainty resulting from ignoring ρ. The CE method on average is the most consistent with q^*. This less costly bias in PD on one hand and the more costly bias in PD_0 and CE_0 on the other hand increases with ρ. Given ρ, all that a higher k does is bring the order quantities under PD_0 and CE_0 closer (as a consequence of reduction in parameter uncertainty). But, the order quantities under PD and CE methods that account for ρ, do not get closer with a higher k. This is because, as mentioned earlier, dependence between experts causes a loss of information that cannot be recovered by increasing k.

In sum, our model PD creates a bias in the order quantity in the less costly direction, whereas PD_0 and CE_0 lead to a bias in the more costly direction. And, CE is the closest to the optimal order quantity under perfect information. But, what does this mean for expected profit under the four methods? We analyze this below.

Figure 2 **Expected order quantities under the four methods (PD, PD_0, CE, CE_0) in the normal model, rescaled with $q^* = 1$, for selected values of CR and k**

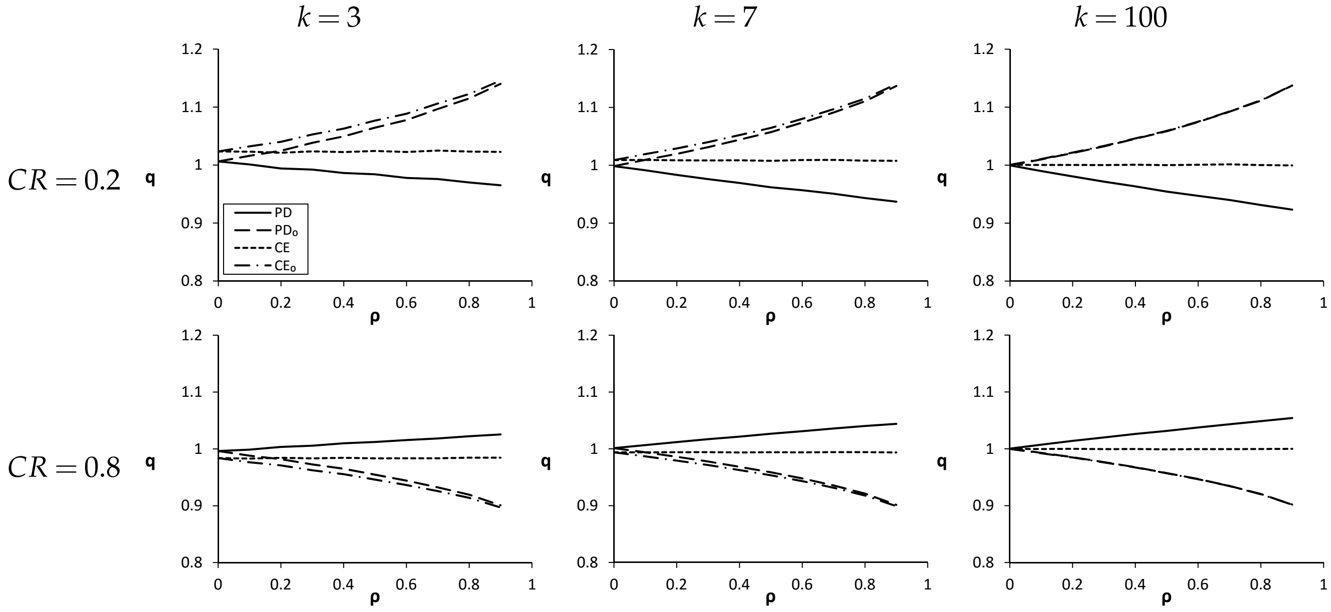
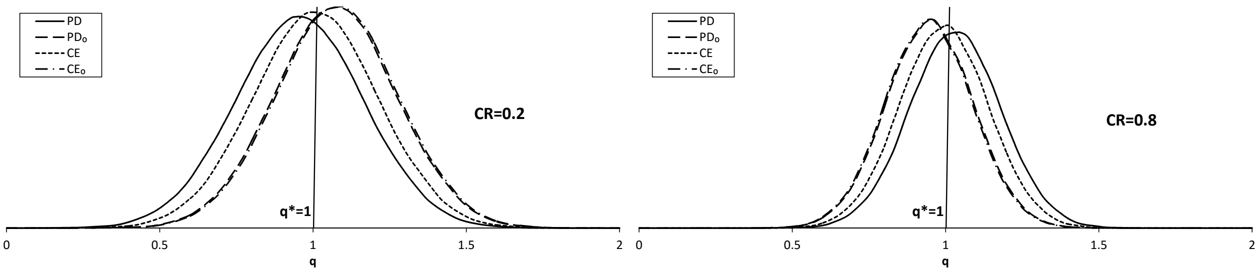


Figure 3 Distribution of q under the four methods (PD, PD_0, CE, CE_0) in the normal model, rescaled with $q^* = 1$, given $\rho = 0.6$, $k = 7$ and $CV = 0.2$, for low and high CR



3.2. Impact of the Demand Estimation Method on the Newsvendor Profit

Here, we investigate the profit distributions associated with each of the estimation methods. For analytical tractability, we first analyze the impact of the estimation method on the newsvendor profit under a simpler setting where σ^2 is known. Later, we show numerically that all the results remain valid also for the case of unknown σ^2.

When demand is normally distributed with unknown μ and known σ^2, with a diffuse prior on μ, the predictive distribution of demand $f(\tilde{y}|x)$ is normal with mean $\bar{x}$ and variance $(1 + 1/k^*)\sigma^2$, where $k^* = k/(1 + (k-1)\rho)$ is the equivalent sample size as before (Clemen and Winkler 1985). The corresponding predictive distribution that ignores ρ is normal with mean $\bar{x}$ and variance $(1 + 1/k)\sigma^2$. With σ^2 known, the CE and CE_0 are the same and involve simply estimating the unknown parameter μ and result in a demand distribution that is normal with mean $\bar{x}$ and variance σ^2. Hence, ρ does not play any role in these two methods.

The optimal order quantities under the different estimation methods are given in Table 4.

Table 4 Optimal order quantities under the four methods (PD, PD_0, CE, CE_0) given x in the normal model with known σ^2

$q_{PD} = \bar{x} + z_{CR}\sigma\sqrt{1 + \frac{1}{k^*}}$
$q_{PD_0} = \bar{x} + z_{CR}\sigma\sqrt{1 + \frac{1}{k}}$
$q_{CE} = q_{CE_0} = \bar{x} + z_{CR}\sigma$

Conditional on μ, $\bar{x} \sim N(\mu, \sigma^2/k^*)$. And, hence, conditional on μ, each of the order quantities in Table 4 are normally distributed with variance σ^2/k^* and

$$E[\tilde{q}_{PD}|\mu] = \mu + z_{CR}\sigma\sqrt{1 + \frac{1}{k^*}}; \quad E[\tilde{q}_{PD_0}|\mu] = \mu + z_{CR}\sigma\sqrt{1 + \frac{1}{k}}; \quad E[\tilde{q}_{CE}|\mu] = \mu + z_{CR}\sigma. \tag{13}$$

Under perfect information about μ and σ^2 (i.e., these are known), the optimal order quantity is $q^* = E[\tilde{q}_{CE}|\mu]$. That is, conditional on μ, the certainty equivalent model in expectation will lead a decision maker to the optimal order quantity. However, as shown next, that expected profit is highest under the *PD* method. That is, $E[\Pi(\tilde{y},\tilde{q}_{PD})] \geq \max\{E[\Pi(\tilde{y},\tilde{q}_{PD_0})],E[\Pi(\tilde{y},\tilde{q}_{CE})]\}$. The following theorem states a more general result.

Theorem 2. *Let $\tilde{y} \sim N(\mu,\sigma^2)$ be the unrealized demand for a newsvendor and $q^* = \mu + z_{CR}\sigma$ be the optimal order quantity under perfect information on the distribution of $\tilde{y}$ (i.e., μ and σ^2 known). Also, let $\tilde{q}_{\epsilon,\tau} \sim N(q^* + \epsilon,\tau^2)$ be a random order quantity. Then*

a) $E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\tau})]$ is decreasing in τ.

b) $E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\tau})]$ is a concave function of ϵ and attains its maximum at $\epsilon^* = z_{CR}\left(\sqrt{\sigma^2+\tau^2}-\sigma\right)$.

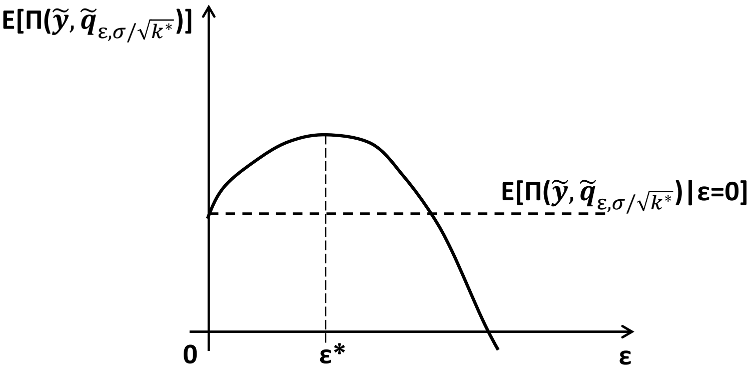
The implication of Theorem 2 for a decision maker is twofold. First, everything else equal, a demand estimation method that leads to an order quantity with a higher variance will result in a lower expected profit. Second, given a variance in the order quantity, a demand estimation method that leads to an order quantity that is "optimal" on average will result in a lower expected profit than a model that errs on the side of caution, provided that the bias is not too severe.

Corollary 2. *a) $E[\Pi(\tilde{y},\tilde{q}_{PD})] \geq E[\Pi(\tilde{y},\tilde{q}_{PD_0})] \geq E[\Pi(\tilde{y},\tilde{q}_{CE})]$.*

b) $E[\Pi(\tilde{y},\tilde{q}_{PD})] = \max_\epsilon E\left[\Pi\left(\tilde{y},\tilde{q}_{\epsilon,\frac{\sigma}{\sqrt{k^*}}}\right)\right]$.

As shown in Corollary 2, among the methods considered, the expected profit is the highest under our *PD* model. Moreover, no other model that leads to an order quantity with equal uncertainty will perform better than the *PD* model in terms of expected profit. Figure 4 illustrates this result for $CR \geq 0.5$, where $\epsilon = 0$ for *CE*, $\epsilon = \epsilon^*$ for *PD*, and $0 < \epsilon < \epsilon^*$ for PD_0.

Figure 4 $E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\sigma/\sqrt{k^*}})]$ **as a function of** ϵ **for** $CR \geq 0.5$



Next, we show with simulations that these results hold also for a normal model with unknown μ and σ^2. For a given *CR*, *CV* =0.2 (WLOG, taking $\mu = 10$ and $\sigma = 2$), k, and ρ, we generate a k-variate observation of $\boldsymbol{x}$ from a multivariate normal and independently generate an observation of $\tilde{y}$ from the underlying normal model ($\mu = 10$ and $\sigma = 2$). For this set of drawn observation, we compute the profit under each of the methods along with profit under q^* (as if μ and σ^2 are known). We do this 100,000 times for each combination of $CR \in \{0.2, 0.8\}$, $CV = 0.2$, $k \in \{3, 7, 100\}$, and $\rho \in \{0, 0.1, ..., 0.9\}$. Figure 5, shows the expected profit, rescaled with $E[\Pi(\tilde{y}, q^*)] = 1$, under the four methods for various levels of the parameters. Figure 6 shows, as an example, the complete distributions of the profit, rescaled after setting $\Pi(\tilde{y}, q^*) = 1$ under the four methods for the case of $CV = 0.2$, $k = 7$, and $\rho = 0.6$ for low CR (0.2) and high CR (0.8). In Figure 5, for all parameter values, the expected profit is highest under PD, followed by CE, then PD_0, and is lowest under CE_0. Given k and $CR \leq 0.5$, the expected profits under PD and CE (methods that account for ρ) are close for low values of ρ but get higher under PD for moderate to large values of ρ. Similarly the expected profits under PD_0 and CE_0 (methods that ignore ρ) are very close, but are less than those under PD and CE with this difference becoming substantially larger at high values of ρ. All that a higher k does is bring PD_0 and CE_0 much closer together. This overall pattern is much less pronounced for $CR \geq 0.5$, where expected profits tend to be much closer under the four methods, but still highest under PD.

For reasons of space, we do not present here the variance of profit for the different parameter values. However, for $CR \leq 0.5$, the variance is lowest under PD, followed by CE, then PD_0, and highest under CE. This is expected as variance of profit is an increasing function of the order quantity (Choi et al. 2008). For $CR \geq 0.5$, the variance of profit is highest under PD, but the differences among the four methods are very small, which is also not surprising as the newsvendor model is much less sensitive to the order quantity relative to when $CR \leq 0.5$. This is somewhat reflected in Figure 6. For $CR = 0.2$, PD_0 and CE_0 have a heavier left tail compared to PD and CE. On the other hand, for $CR = 0.8$, the same left tail is much more similar under all four methods.

To summarize, we show that the four methods of demand estimation can have a significant impact on the optimal order quantity and profit in a newsvendor setting. This impact is the highest for low critical ratios. In particular, we have shown that our model PD yields the highest (lowest) order quantity of all the models under consideration for high (low) CRs. In other words, our model is biased in the direction of the less costly decision (i.e., underestimation of demand with low CR and overestimation of demand with high CR). Further, the expected profit is the highest under our model. For low CRs, this increase in expected profit can exceed 20% compared to the other methods discussed. For high CRs, the impact is more muted, and the increase in expected profit under our model compared to the other methods is generally less than 5%. This is intuitive, because the newsvendor model is relatively insensitive to the order quantity for high

Figure 5 **Expected profit under the four methods (PD, PD_0, CE, CE_0) in the normal model, rescaled with $E[\Pi(q^*)] = 1$, given $CV = 0.2$, for selected values of CR and k**

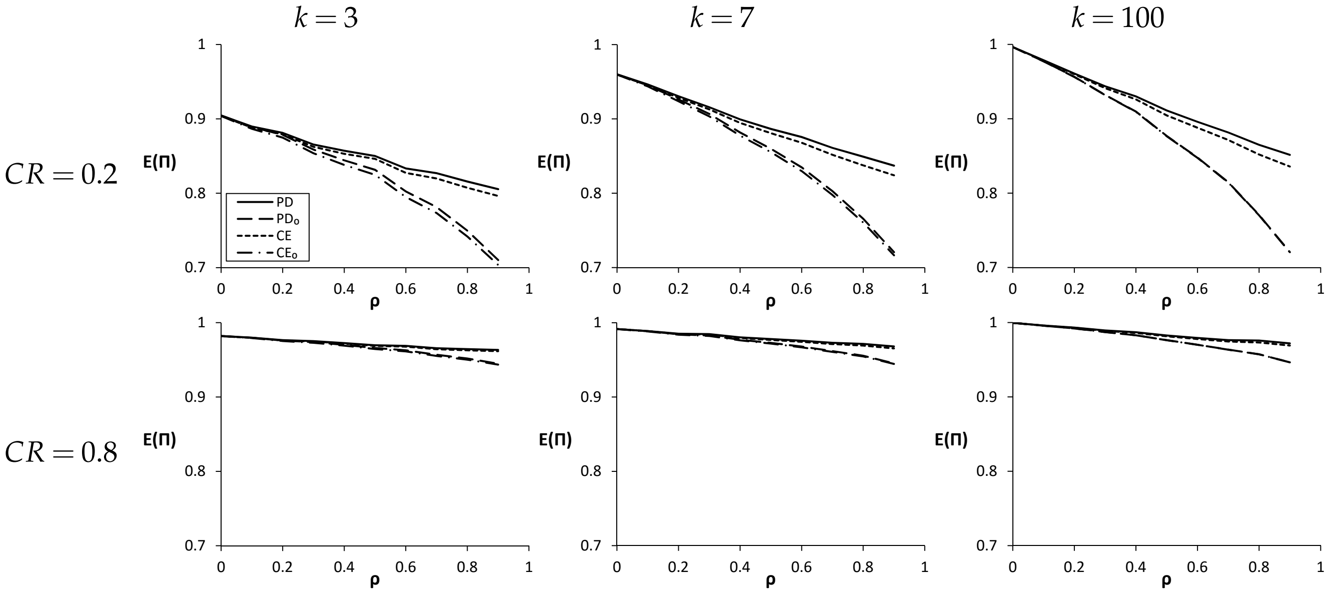
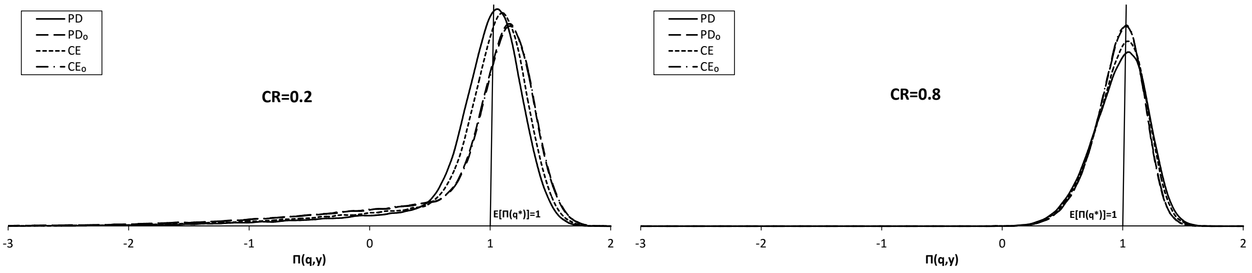


Figure 6 **Distribution of $\Pi(q, y)$ under the four methods (PD, PD_0, CE, CE_0) in the normal model, rescaled with $E[\Pi(q^*)] = 1$, given $\rho = 0.6$, $k = 7$ and $CV = 0.2$, for low and high CR**



CRs for normally distributed demand (for example, simply ordering the mean demand will yield very similar profits as optimal order quantities under the four methods).

Ignoring ρ or the parameter uncertainty leads to lower expected profits. Ignoring ρ appears to be a more costly mistake than ignoring parameter uncertainty, with expected profit being higher under CE than under PD_0. Not only is the expected profit highest under our model, but the variance of profit is the lowest for $CR \leq 0.5$. For $CR \geq 0.5$, the variance under PD is the highest, but the differences across the four methods are very small, in fact close to negligible.

We next investigate some extensions of the normal model in §2.

4. Extensions

In the model in §2, we have assumed that the pairwise correlation between experts is the same and known. In this section, we introduce uncertainty on the common correlation ρ and also heterogeneity in the pairwise correlations among the experts, to check for robustness of the augmentation factor δ shown in Table 2. We then extend our normal model to a lognormal model.

4.1. Uncertainty and Heterogeneity about ρ in the Normal Model

We first explore uncertainty on the common ρ. Recall that in our model, δ depends only on ρ and k. Given k, δ increases with ρ, indicating greater uncertainty about $\tilde{y}|\boldsymbol{x}$ due to greater dependence between the experts. On the other hand, given ρ, δ decreases with k, reflecting lower uncertainty about $\tilde{y}|\boldsymbol{x}$ due to reduced parameter uncertainty resulting from a higher number of experts. For a given k, we take 10,000 independent draws of ρ from a beta distribution, $f(\rho) \propto \rho^{\alpha-1}(1-\rho)^{\beta-1}$, with $\alpha, \beta > 0, E(\rho) = \alpha/(\alpha+\beta)$, and $Var(\rho) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$, and calculate the average δ. We repeat this for k=3, 5, and 10, and for a variety of values for the parameters of the beta distribution. Keeping $E(\rho)$ constant at 0.1, 0.2, ..., 0.9, we vary $\alpha+\beta$ from 100 (low uncertainty) to 50 (medium uncertainty) to 10 (high uncertainty). Table 5 below shows the average δ for different values of k and $E(\rho)$ with varying levels of uncertainty, along with the case of known ρ (no uncertainty).

Table 5 The impact of uncertainty about ρ on δ:
average δ for selected values of k and $E(\rho)$ with varying levels of uncertainty about ρ

	k and levels of uncertainty about ρ											
	$k=3$				$k=5$				$k=10$			
$E(\rho)$	None	Low	Medium	High	None	Low	Medium	High	None	Low	Medium	High
0.1	1.76	1.76	1.77	1.77	1.38	1.38	1.38	1.38	1.22	1.22	1.22	1.23
0.2	1.91	1.92	1.92	1.94	1.51	1.51	1.51	1.52	1.34	1.34	1.35	1.36
0.3	2.09	2.10	2.10	2.14	1.66	1.66	1.66	1.69	1.48	1.49	1.49	1.52
0.4	2.31	2.32	2.32	2.39	1.84	1.84	1.85	1.89	1.65	1.66	1.66	1.71
0.5	2.58	2.59	2.61	2.71	2.07	2.07	2.08	2.17	1.87	1.88	1.89	1.96
0.6	2.94	2.97	2.98	3.14	2.37	2.38	2.40	2.55	2.15	2.16	2.18	2.30
0.7	3.46	3.50	3.53	3.84	2.80	2.82	2.85	3.12	2.55	2.57	2.60	2.83
0.8	4.32	4.40	4.45	5.24	3.50	3.55	3.63	4.26	3.20	3.26	3.30	3.90
0.9	6.22	6.44	6.69	10.76	5.06	5.26	5.49	8.47	4.64	4.79	4.99	8.13

First note that, given k, δ under no uncertainty is the lowest. It makes sense that any uncertainty about ρ is yet another addition to the parameter uncertainty and should increase δ. Further, for $E(\rho) \leq 0.6$, δ does not vary much at all with increasing uncertainty about ρ. Even for $E(\rho) \geq 0.7$, the impact is less than remarkable. For example, with $E(\rho) = 0.7$ the increase in δ from no uncertainty to high uncertainty is 3.46 to 3.84 for $k = 3$, 2.8 to 3.12 for $k = 5$, and 2.55 to 2.83 for $k = 10$. It is only in the extreme case of high uncertainty and very high $E(\rho)$ at 0.9 that the jump in δ is substantial. In sum, it appears that δ is reasonably robust with respect to uncertainty about ρ. In any case, the no uncertainty case (i.e., known ρ) provides a lower bound on δ. And, this lower bound could be a reasonable value of δ to use in most practical situations even if there is uncertainty about ρ.

Next, we explore heterogeneity among the experts in terms of the pairwise correlations. For such a case, our model with a common ρ no longer holds. In Appendix B, we show a more general model, where $\boldsymbol{x} = (x_1, ..., x_k)'$ follows a multivariate normal with mean vector $\boldsymbol{\mu} = \mu \boldsymbol{e}$, where $\boldsymbol{e} = (1, ..., 1)'$ is a $k \times 1$ column vector, and $k \times k$ positive definite covariance matrix $\boldsymbol{\Sigma}$ with diagonal elements σ^2 and off-diagonal elements $\rho_{ij}\sigma^2$, with ρ_{ij} known. Both $E(\tilde{y}|\boldsymbol{x})$ and $Var(\tilde{y}|\boldsymbol{x})$ are weighted mixtures of all the ρ_{ij}s and the observed x. And, the augmentation factor δ depends not only of k and ρ_{ij}s but also on the observed $\boldsymbol{x}$.

To investigate the impact of heterogeneity in ρ_{ij}s, for a given group of k experts, we generate each pairwise correlation ρ_{ij} independently from a beta distribution with the same parameters as earlier. Here, the mean of the beta distribution is the average ρ_{ij} among the k experts, and the variance is the level of heterogeneity (higher variance indicating higher heterogeneity). If the draws result in a covariance matrix that is not positive definite, then that draw is discarded and another is taken until there are 10,000 draws in all. For each draw of a positive definite covariance matrix, we generate a k-variate forecast from a multivariate normal with mean vector 0 and a covariance matrix based on the drawn observation of a set of ρ_{ij}s and with $\sigma^2 = 1$ (the choice of 0 mean vector and $\sigma^2 = 1$ is WLOG, as shown in Appendix C). We then compute δ as the ratio of $\sqrt{Var(\tilde{y}|\boldsymbol{x})}$ with respect to the observed s. The proportion of positive definite covariance matrices is small for large k and for high levels of heterogeneity with high $E(\rho)$. So, we generate simulations for k = 3, 5, and 10, and holding *average* ρ_{ij}=$\alpha/(\alpha+\beta)$ constant at 0.1, 0.2, ..., 0.7, we vary $\alpha + \beta$ (the level of heterogeneity) from 100 (low) to 50 (medium), along with the corresponding case where the average ρ_{ij} is the common ρ (none). For what we label medium, the degree of heterogeneity is still reasonably high. For example, with average ρ_{ij} = 0.5, the standard deviation of beta distribution is 0.07 for medium heterogeneity and 0.05 for low heterogeneity. Table 6 below shows the average δ for these cases.

Once again, the overall results are similar to the case of known and common ρ (shown in Table 2). Given k, a higher average ρ_{ij} leads to a higher δ, and given average ρ_{ij} a higher k leads to lower δ. What is more reassuring is that, at least for the results we obtained, heterogeneity has none to low effect on δ.

In sum, the results indicate that the δ in our model with common and known ρ provides a lower bound for cases where ρ is common but uncertain and for cases where ρ equals the average of ρ_{ij}s. The level of uncertainty on the common ρ matters substantially only for the extreme cases of high uncertainty and very high $E(\rho)$, whereas δ is very robust with respect to heterogeneity in pairwise correlations. This provides some further encouragement for our model with a common and known ρ.

Table 6 The impact of heterogeneity in pairwise correlations on δ:
average δ for selected values of k and average ρ_{ij} with varying levels of heterogeneity in ρ_{ij}s

	k and levels of heterogeneity in ρ_{ij}s								
	$k=3$			$k=5$			$k=10$		
Average ρ_{ij}	None	Low	Medium	None	Low	Medium	None	Low	Medium
0.1	1.76	1.77	1.76	1.38	1.38	1.38	1.22	1.22	1.22
0.2	1.91	1.92	1.92	1.51	1.51	1.51	1.34	1.34	1.34
0.3	2.09	2.10	2.10	1.66	1.66	1.66	1.48	1.49	1.49
0.4	2.31	2.32	2.32	1.84	1.84	1.85	1.65	1.66	1.66
0.5	2.58	2.59	2.60	2.07	2.07	2.08	1.87	1.88	1.89
0.6	2.94	2.96	2.97	2.37	2.38	2.39	2.15	2.17	2.18
0.7	3.46	3.49	3.52	2.80	2.82	2.85	2.55	2.57	2.58

4.2. A Lognormal Model

In §2.1, we assume that the distribution of $\tilde{y}$ given μ and σ^2 is normal. This is often a reasonable approximation for many real-life settings. However, in some instances, the decision maker might consider the distribution above to be skewed and/or bounded from below with a long right tail, such as demand for a new product that has a small chance but a large potential for blockbuster sales. In such cases, a lognormal distribution might be more appropriate than a normal. In that spirit, we modify our model in §2.1 such that $\ln \tilde{y} \sim N(\mu, \sigma^2)$ and $\ln x_1, ..., \ln x_k$ follow a multivariate normal with mean vector $\boldsymbol{\mu} = \mu \boldsymbol{e}$, where $\boldsymbol{e} = (1, ..., 1)'$ is a $k \times 1$ column vector, and $k \times k$ positive definite covariance matrix $\boldsymbol{\Sigma}$ with diagonal elements σ^2 and off-diagonal elements $\rho_L \sigma^2$, where $\rho_L = Corr[\ln x_i, \ln x_j | \theta], i \neq j$, which implies that $\rho = Corr[x_i, x_j | \theta] = \frac{e^{\rho_L \sigma^2} - 1}{e^{\sigma^2} - 1}$ (Johnson et al. 2002). Note that $\rho_L = 0$ when $\rho = 0$, $\rho_L = 1$ when $\rho = 1$, and they have a monotonically increasing relationship between 0 and 1, where ρ_L depends also on σ (and hence the CV of the distribution).

In our normal model, the sufficient statistics are $\overline{x}$ and s^2, the sample mean and the sample variance of the k point estimates, such that $f(\tilde{y}|\boldsymbol{x}) = f(\tilde{y}|\overline{x}, s^2)$ is a t distribution. Using the same results, in our lognormal model, the sufficient statistics are $\overline{x}_t = \frac{1}{k}\sum_{i=1}^{k} \ln x_i$ and $s_t^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\ln x_i - \overline{x}_t)^2$. And, $f(\ln \tilde{y}|\boldsymbol{x}) = f(\ln \tilde{y}|\overline{x}_t, s_t^2)$ is a t distribution with the same parameters as in the normal model except for $\overline{x}_t$ and s_t^2 replacing $\overline{x}$ and s^2, respectively. The distribution for $\tilde{y}|\boldsymbol{x}$ is then a *log t* distribution which is very heavy tailed with undefined moments. However, it is still possible to construct prediction intervals for $\tilde{y}|\boldsymbol{x}$ by simply taking the exponential transforms of the corresponding quantiles of the underlying t distribution for $f(\ln \tilde{y}|\boldsymbol{x})$. Table 7 below summarizes the lognormal model under the four methods (PD, PD_0, CE, CE_0) defined in §2.1.1, and provides for each of those the distribution of $\ln \tilde{y}|\boldsymbol{x}$ and the corresponding $100(1-\gamma)\%$ prediction interval, $0 \leq \gamma \leq 1$, for $\tilde{y}|\boldsymbol{x}$.

Table 7 A summary of the four methods (PD, PD_0, CE, CE_0) in the lognormal model

	$f(\ln\tilde{y}\mid\boldsymbol{x})$	A $100(1-\gamma)\%$ Prediction Interval for $\tilde{y}\mid\boldsymbol{x}$
PD	$t_k(\overline{x}_t, \sqrt{\frac{k-1}{k}(\frac{1+\rho_L}{1-\rho_L}+\frac{1}{k})}s_t)$	$exp(\overline{x}_t \pm t_{1-\frac{\gamma}{2},k}\sqrt{\frac{k-1}{k}(\frac{1+\rho_L}{1-\rho_L}+\frac{1}{k})}s_t)$
PD_0	$t_k(\overline{x}_t, \sqrt{\frac{k-1}{k}(1+\frac{1}{k})}s_t)$	$exp(\overline{x}_t \pm t_{1-\frac{\gamma}{2},k}\sqrt{\frac{k-1}{k}(1+\frac{1}{k})}s_t)$
CE	$N(\overline{x}_t, \frac{s_t^2}{1-\rho_L})$	$exp(\overline{x}_t \pm z_{1-\frac{\gamma}{2}}\sqrt{\frac{1}{1-\rho_L}}s_t)$
CE_0	$N(\overline{x}_t, s_t^2)$	$exp(\overline{x}_t \pm z_{1-\frac{\gamma}{2}}s_t)$

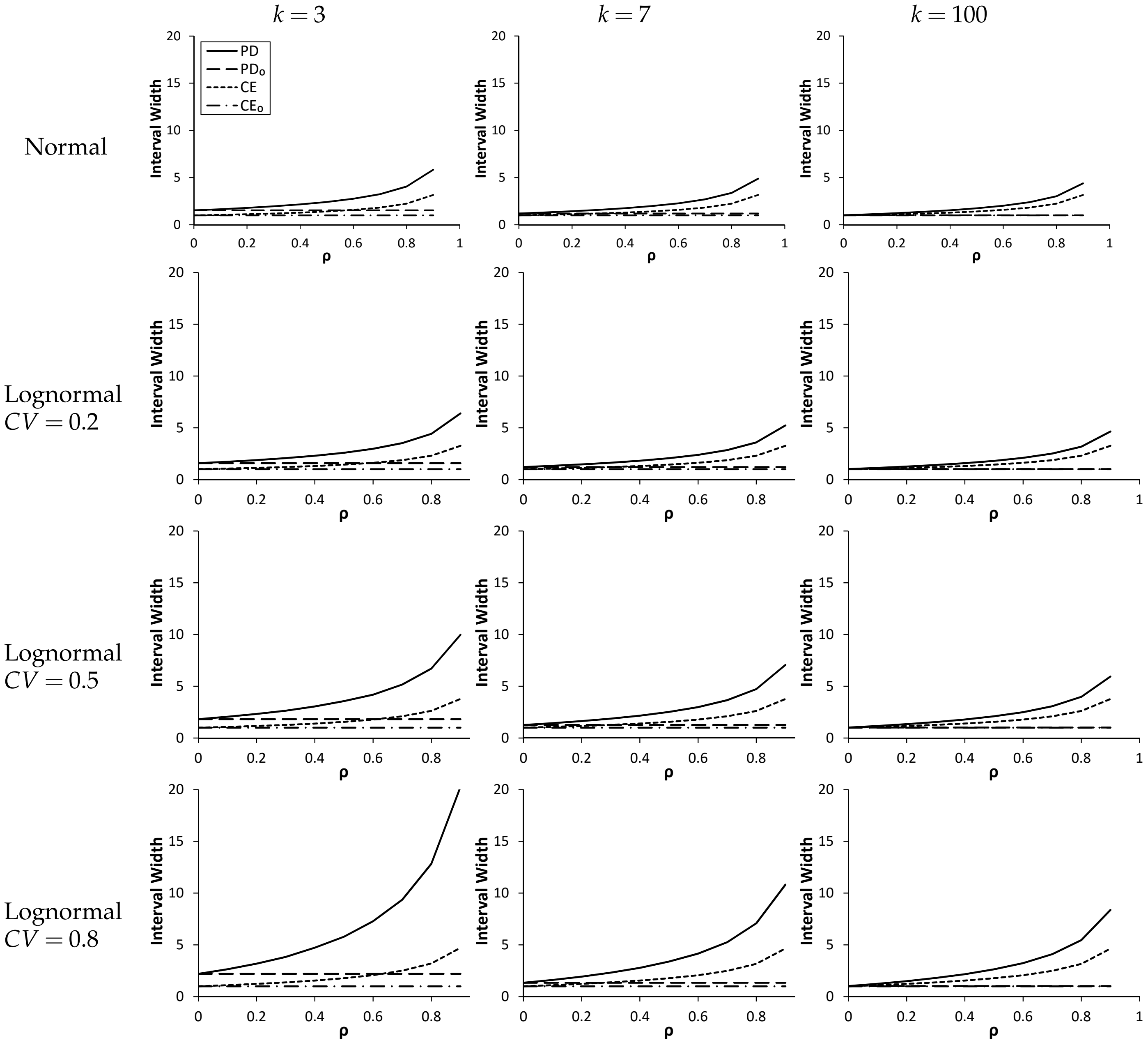
Note that the prediction intervals are stated in terms of $\rho_L = Corr[\ln x_i, \ln x_j|\theta]$. It is easy to see that the rank order of the widths of these prediction intervals under the four methods is the same as in our normal model. The prediction interval for PD (CE_0) are the most wide (narrow). Ignoring parameter uncertainty and/or the dependence between the experts creates spurious accuracy in terms of an unrealistically narrow prediction interval for $\tilde{y}|\boldsymbol{x}$.

We also do some simulations to further explore the lognormal model. WLOG, we use a zero vector for $\boldsymbol{\mu}$ (See Appendix D). For a given k, ρ, and CV, we generate a k-variate observation of $\boldsymbol{x}$ from the corresponding multivariate lognormal. For each observation, we compute the width of a 95% prediction interval for $f(\tilde{y}|\boldsymbol{x})$ under each of the four methods, rescaled with the width under CE_0 set equal to 1. A comparison of the four widths indicates the impact of ignoring the parameter uncertainty or ρ or both in a 95% prediction interval. For each combination of k=3, 7, and 100, and CV= 0.2, 0.5 and 0.8, we repeat this 10,000 times and compute the average rescaled widths under the four methods. Figure 7 shows these results, along with the corresponding average rescaled widths under the four methods in the normal model (the first row of three graphs) which do not depend on CV.

The overall results are consistent with the normal model in terms of the rank orders of the widths, except that CV plays a salient role in the lognormal model. Given k and $CV = 0.2$, the widths of the intervals for the normal and the lognormal are similar. However, given k, as CV increases, the width under PD in the lognormal model gets substantially higher. The main implication we see from Figure 7 is that our PD approach in the lognormal model is even more crucial for a higher CV.

To summarize the normal and lognormal models, we propose a predictive distributions for $\tilde{y}|x$ that incorporates the parameter uncertainty and the dependence between the experts. Ignoring one or both can lead to serious underestimation of uncertainty. If there is dependence between the experts, there is a loss of information that can not be compensated by simply increasing the number of experts, even to an extreme. Any method that ignores the dependence

Figure 7 **Average widths of 95% prediction intervals under the four methods (PD, PD_0, CE, CE_0) in the normal and lognormal models, rescaled with average width under CE_0 =1, for selected values of CV and k**



necessarily underestimates the uncertainty about $\tilde{y}|x$. Similarly, a certainty equivalent approach even if it accounts for dependence between experts underestimates uncertainty about $\tilde{y}|x$. Such underestimation gets worse for a higher ρ.

In the normal model, the predictive distribution for $\tilde{y}|x$ is a t distribution with degrees of freedom k, location $\overline{x}$, and scale δs, which implies that $E(\tilde{y}|x) = \overline{x}$ and $Var(\tilde{y}|x) = (\delta s\sqrt{k/(k-2)})^2$, where $\delta = \sqrt{\frac{k-1}{k}(\frac{1+\rho}{1-\rho} + \frac{1}{k})}$. Similarly, in the lognormal model, the predictive distribution for $\ln \tilde{y}|x$ is a t distribution with degrees of freedom k, location $\overline{x}_t$, and scale $\delta_L s_t$, which implies that

$E(\ln \tilde{y}|\boldsymbol{x}) = \overline{x}_t$ and $Var(\ln \tilde{y}|\boldsymbol{x}) = (\delta_L s_t \sqrt{k/(k-2)})^2$ where δ_L is the same as in the normal model but with ρ_L replacing ρ, i.e., $\delta_L = \sqrt{\frac{k-1}{k}(\frac{1+\rho_L}{1-\rho_L} + \frac{1}{k})}$. The factor δ (or δ_L) adjusts the observed sample standard deviation of the k point estimates upwards for any given dependence between the experts and for any remaining parameter uncertainty given the k point estimates. These predictive distributions are fairly easy to use, and in that sense practically viable.

5. Summary and Discussion

We develop a parsimonious approach for generating a probability distribution for a variable of interest based on point forecasts provided by experts. Our approach allows for the possibility of point forecasts to be correlated and admits parameter uncertainty given the forecasts. In keeping with the extensive empirical findings on combining forecasts, we use an equal-weights model for experts, i.e., all experts are treated equally in terms of their accuracy and with a common correlation between their forecasts. The resulting predictive distribution shows that ignoring either the parameter uncertainty or the dependence between experts can lead to much spurious accuracy in terms of an unrealistically narrow distribution for the variable of interest. Further, we provide a rationale for the augmentation factor used often in the operations management literature, as a simple scalar factor that is needed to equate the standard deviation of the predictive distribution to the observed standard deviation of the point forecasts. This augmentation factor, which is always greater than 1, depends on the common correlation between the experts and on the number of experts. Given a number of experts, a higher correlation leads to a higher augmentation factor, indicating greater uncertainty for the quantity of interest. This is consistent with earlier findings in the Bayesian literature that greater dependence among experts causes a greater loss of information. On the other hand, given a correlation between the experts, a higher number of experts reduces the augmentation factor, indicating lower parameter uncertainty. However, loss of information due to dependence between experts cannot be overcome by simply increasing the number of experts, even to an extreme. For example, given $\rho = 0.8$, even with 100 point forecasts, the augmentation factor remains at about 3. We compare our model with other methods that ignore either the dependence between experts or the parameter uncertainty, or both, illustrating potentially serious consequences in terms of underestimating the uncertainty on the variable of interest.

We illustrate the same consequences in a decision making context of a newsvendor setting. Our model, when compared to other methods that ignore dependence or parameter uncertainty, or both, leads to an order quantity that on average is smaller (larger) for $CR \leq 0.5$ ($CR \geq 0.5$), but has higher variance. However, in the same comparison, our model leads to the highest expected profit. In fact, we show that a method that ignores dependence and parameter uncertainty leads to on average an order quantity that is optimal under perfect information on the distribution parameters, but yet yields lower expected profit compared to our model. This is because, given the uncertainty in the demand distribution after observing the point forecasts, our model errs on the side of caution, whereas a method that ignores dependence and parameter uncertainty retains much spurious accuracy (in terms of a distribution that is tighter than it should be) and hence a greater chance of a costly mistake (in terms of the order quantity). The increase in expected profit in our model can exceed 20%

In our base model, we assume that the common correlation between the experts is known. However, we investigate the robustness of our model in terms of uncertainty on the common correlation between experts and heterogeneity in pairwise correlations between experts. The augmentation factor from our model is fairly robust in these respects. We further extend our model under normality to the case where the quantity of interest and the point forecasts have a lognormal distribution. The results under normality only get exacerbated in the lognormal case, i.e., the consequences of ignoring dependence or parameter uncertainty lead to even more unrealistically narrower distributions for the quantity of interest.

We feel that our model is not only more accurate but practically viable. A step-by-step practical implementation of our model would require the decision maker to do the following. First, the decision maker must make a judgment whether the underlying distribution is normal or lognormal. These two distributions can be reasonable approximations in a wide variety of real-life situations. However, if it is strongly felt that the underlying distribution is other than normal or lognormal, a model akin to our approach can be easily developed. If the decision maker has any prior information on the distribution parameters, that must then be reflected in a prior distribution. Else, a diffuse prior can be used, allowing the data speak for themselves. The next step requires estimation of the common correlation between a given set of experts. A full-blown discussion on this is beyond the scope of the paper. However, there exist extensive literature on such estimation with or without past data (Gokhale and James 1982, Clemen et al. 2000, Meyer and Booker 2001, to name only a few). Thinking of the common correlation as the proportion of total information that is common across experts (Lichtendahl Jr et al. 2013, Winkler 1981) for example, could be a further aid in such estimation. It is worthwhile to mention again that our model is fairly robust to uncertainty about the common correlation and to heterogeneity in pairwise correlations among experts. In any case, the mean correlation in case of uncertainty or heterogeneity provides at the very least in our model a lower bound on the uncertainty about the variable of interest. Clearly, simply ignoring the dependence among experts is not a good option. Given the number of experts and an estimated common correlation between them, computing the appropriate augmentation factor is straightforward. What then remains is obtaining the point forecasts and computing the relevant sufficient statistics.

Appendix

A. Proofs

Proof of Lemma 1:

Proof:

$$
\begin{aligned}
l(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \frac{\lambda^{\frac{k}{2}}}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}_\rho|}} exp\left(-\frac{\lambda}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_\rho^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&\propto \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_\rho^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda}{2}(\boldsymbol{x}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x} - 2\mu e'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x} + \mu^2 e'\boldsymbol{\Sigma}_\rho^{-1}e)\right) \\
&= \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda e'\boldsymbol{\Sigma}_\rho^{-1}e}{2}(\frac{\boldsymbol{x}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e} - 2\mu\frac{e'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e} + \mu^2)\right) \\
&= \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda e'\boldsymbol{\Sigma}_\rho^{-1}e}{2}(\mu - \frac{e'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e})^2\right) exp\left(-\frac{\lambda e'\boldsymbol{\Sigma}_\rho^{-1}e}{2}(\frac{\boldsymbol{x}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e} - (\frac{e'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e})^2)\right) \\
&= \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda k^*}{2}(\mu-\overline{x})^2\right) exp\left(-\frac{\lambda}{2}(k-1)s^{*2}\right).
\end{aligned}
$$

It is easy to verify that $e'\boldsymbol{\Sigma}_\rho^{-1}e = k^*$, with $k^* = k/(1+(k-1)\rho)$; $\frac{e'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e} = \overline{x}$, where $\overline{x} = (1/k)\sum_{i=1}^k x_i$; and $\frac{\boldsymbol{x}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e} - \left(\frac{e'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{x}}{e'\boldsymbol{\Sigma}_\rho^{-1}e}\right)^2 = (k-1)s^{*2}$, where $s^{*2} = s^2/(1-\rho)$ and $s^2 = \sum_{i=1}^k (x_i-\overline{x})^2/(k-1)$. ■

Proof of Theorem 1:

Proof:

$$
\begin{aligned}
f(\mu,\lambda|\boldsymbol{x}) &\propto NG(\mu,\lambda|\mu_0,n_\mu,v_0,n_v)l(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\propto \lambda^{\frac{1}{2}} exp\left(-\frac{n_\mu\lambda(\mu-\mu_0)^2}{2}\right)\lambda^{\frac{n_v}{2}-1} exp\left(-\frac{n_v v_0}{2}\lambda\right)\lambda^{\frac{k}{2}} exp\left(-\frac{\lambda k^*}{2}(\mu-\overline{x})^2\right) exp\left(-\frac{\lambda}{2}(k-1)s^{*2}\right) \\
&= \lambda^{\frac{1}{2}} exp\left(-\frac{\lambda}{2}\left(n_\mu(\mu-\mu_0)^2 + k^*(\mu-\overline{x})^2\right)\right)\lambda^{\frac{n_v}{2}+\frac{k}{2}-1} exp\left(-\left(\frac{n_v v_0}{2} + \frac{(k-1)s^{*2}}{2}\right)\lambda\right).
\end{aligned}
$$

Based on the identity $n_\mu(\mu-\mu_0)^2 + k^*(\mu-\overline{x})^2 = (n_\mu+k^*)(\mu-\mu^*)^2 + \frac{n_\mu k^*(\overline{x}-\mu_0)^2}{n_\mu+k^*}$, where $\mu^* = \frac{n_\mu\mu_0+k^*\overline{x}}{n_\mu+k^*}$,

$$
\begin{aligned}
f(\mu,\lambda|\boldsymbol{x}) &\propto \lambda^{\frac{1}{2}} exp\left(-\frac{\lambda}{2}(n_\mu+k^*)(\mu-\mu^*)^2\right)\lambda^{\frac{n_v}{2}+\frac{k}{2}-1} exp\left(-\left(\frac{n_v v_0}{2} + \frac{(k-1)s^{*2}}{2} + \frac{n_\mu k^*(\overline{x}-\mu_0)^2}{2(n_\mu+k^*)}\right)\lambda\right) \\
&\propto NG(\mu,\lambda|\mu^*,n_\mu^*,v^*,n_v^*).
\end{aligned}
$$

where $n_\mu^* = n_\mu + k^*$, $\mu^* = \frac{n_\mu\mu_0+k^*\overline{x}}{n_\mu^*}$, $n_v^* = n_v + k$ and $v^* = \frac{1}{n_v^*}\left(n_v v_0 + (k-1)s^{*2} + \frac{n_\mu k^*}{n_\mu+k^*}(\overline{x}-\mu_0)^2\right)$ ■

Proof of Corollary 1:

Proof: The predictive distribution $\tilde{y}|\boldsymbol{x}$

$$
\begin{aligned}
f(\tilde{y}|\boldsymbol{x}) &= \int_\lambda \int_\mu f(y|\mu,\lambda) f(\mu,\lambda|\boldsymbol{x}) d\mu d\lambda \\
&\propto \int_\lambda \int_\mu \lambda^{\frac{1}{2}} exp\left(-\frac{\lambda}{2}(y-\mu)^2\right) \lambda^{\frac{1}{2}} exp\left(-\frac{\lambda}{2} n_\mu^*(\mu-\mu^*)^2\right) \lambda^{\frac{n_v^*}{2}-1} exp\left(-\frac{n_v^* v^*}{2}\lambda\right) d\mu d\lambda \\
&= \int_\lambda \int_\mu exp\left(-\frac{\lambda}{2}\left((y-\mu)^2 + n_\mu^*(\mu-\mu^*)^2\right)\right) \lambda^{\frac{n_v^*}{2}} exp\left(-\frac{n_v^* v^*}{2}\lambda\right) d\mu d\lambda.
\end{aligned}
$$

Using the identity $(y-\mu)^2 + n_\mu^*(\mu-\mu^*)^2 = (1+n_\mu^*)(\mu-\mu_n^*)^2 + \frac{n_\mu^*(y-\mu^*)^2}{n_\mu^*+1}$ where $\mu_n^* = \frac{y+n_\mu^*\mu^*}{n_\mu^*+1}$,

$$
\begin{aligned}
f(\tilde{y}|\boldsymbol{x}) &\propto \int_\lambda \int_\mu exp\left(-\frac{\lambda}{2}(1+n_\mu^*)(\mu-\mu_n^*)^2\right) exp\left(-\frac{\lambda}{2}\frac{n_\mu^*(y-\mu^*)^2}{n_\mu^*+1}\right) \lambda^{\frac{n_v^*}{2}} exp\left(-\frac{n_v^* v^*}{2}\lambda\right) d\mu d\lambda \\
&\propto \int_\lambda exp\left(-\frac{\lambda}{2}\frac{n_\mu^*(y-\mu^*)^2}{n_\mu^*+1}\right) \lambda^{\frac{n_v^*}{2}} exp\left(-\frac{n_v^* v^*}{2}\lambda\right) \lambda^{-\frac{1}{2}} d\lambda \\
&= \int_\lambda exp\left(-\left(\frac{n_v^* v^*}{2} + \frac{n_\mu^*(y-\mu^*)^2}{2(n_\mu^*+1)}\right)\lambda\right) \lambda^{\frac{n_v^*}{2}-\frac{1}{2}} d\lambda.
\end{aligned}
$$

We recognize this is a Gamma distribution with shape parameter $a = \frac{n_v^*}{2} + \frac{1}{2}$ and rate parameter $b = \frac{n_v^* v^*}{2} + \frac{n_\mu^*(y-\mu^*)^2}{2(n_\mu^*+1)}$.

$$
\begin{aligned}
f(\tilde{y}|\boldsymbol{x}) &\propto b^{-a} \\
&= \left(\frac{n_v^* v^*}{2} + \frac{n_\mu^*(y-\mu^*)^2}{2(n_\mu^*+1)}\right)^{-(\frac{n_v^*}{2}+\frac{1}{2})} \\
&\propto \left(1 + \frac{1}{n_v^*}\frac{(y-\mu^*)^2}{\frac{n_\mu^*+1}{n_\mu^*}v^*}\right)^{-(\frac{n_v^*+1}{2})}.
\end{aligned}
$$

The predictive distribution for $\tilde{y}|\boldsymbol{x}$ is a t distribution with degrees of freedom n_v^*, location parameter μ^*, and scale parameter $\sqrt{(n_\mu^*+1)v^*/n_\mu^*}$. ■

PROOF OF PROPOSITION 1:

Proof: For $CR \leq 0.5$, $\tilde{q}_{PD} \leq \min\left\{\tilde{q}_{PD_0}, \tilde{q}_{CE}, \tilde{q}_{CE_0}\right\}$ for any given $\overline{x}$ and s. Hence, the inequality holds for the expected values as well. Similar argument holds for $CR \geq 0.5$.

We now show that $Var[\tilde{q}_{PD}] \geq Var[\tilde{q}_{PD_0}]$. We make use of a known result that $\overline{x}$ and s^2 are independently distributed when x_is follow a multivariate normal distribution with identical pairwise ρs (Rao 1973, p. 196-197). It then follows that $\overline{x}$ and s are independently distributed, i.e., $Cov[\overline{x}, s] = 0$.

Hence,

$$
\begin{aligned}
Var[\tilde{q}_{PD}] &= Var[\overline{x}] + t_{CR,k}^2 \frac{k-1}{k}\left(\frac{1+\rho}{1-\rho} + \frac{1}{k}\right) Var[s], \text{and} \\
Var[\tilde{q}_{PD_0}] &= Var[\overline{x}] + t_{CR,k}^2 \frac{k-1}{k}\left(1 + \frac{1}{k}\right) Var[s].
\end{aligned}
$$

Since $\frac{k-1}{k}\left(\frac{1+\rho}{1-\rho} + \frac{1}{k}\right) \geq \frac{k-1}{k}\left(1 + \frac{1}{k}\right)$, it directly follows that $Var[\tilde{q}_{PD}] \geq Var[\tilde{q}_{PD_0}]$. The same argument holds for $Var[\tilde{q}_{PD}] \geq Var[\tilde{q}_{CE}]$ and $Var[\tilde{q}_{PD}] \geq Var[\tilde{q}_{CE_0}]$. ■

PROOF OF THEOREM 2:

Proof: a) Given an order quantity, the expected profit is given by:

$$E[\Pi(\tilde{y}|\tilde{q}_{\epsilon,\tau}=q]=(p-v)\mu-(c-v)q-\sigma L(z)(p-v), \tag{A.1}$$

where $z=\frac{q-\mu}{\sigma}$ and $L(z)=\int_z^\infty(x-z)\,\mathrm{d}\Phi(x)$. Then,

$$E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\tau})]=(p-v)\mu-(c-v)(q^*+\epsilon)-\sigma(p-v)\int_{-\infty}^{\infty}L(z)\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{(q-q^*-\epsilon)^2}{2\tau^2}}\,\mathrm{d}q. \tag{A.2}$$

And,

$$\frac{\partial E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\tau})]}{\partial\tau}=\frac{\sigma}{\tau}(p-v)\int_{-\infty}^{\infty}L(z)\frac{e^{-\frac{(q-q^*-\epsilon)^2}{2\tau^2}}}{\tau\sqrt{2\pi}}dq-\sigma(p-v)\int_{-\infty}^{\infty}L(z)\frac{(q-q^*-\epsilon)^2}{\tau^3}\frac{e^{-\frac{(q-q^*-\epsilon)^2}{2\tau^2}}}{\tau\sqrt{2\pi}}\,\mathrm{d}q.$$

Integrating by parts twice the second term in the expression, we get:

$$\frac{\partial E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\tau})]}{\partial\tau}=-\tau\sigma(p-v)\int_{-\infty}^{\infty}\frac{\partial^2 L(z)}{\partial q^2}\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{(q-q^*-\epsilon)^2}{2\tau^2}}\,\mathrm{d}q.$$

But $L(z)$ is a convex decreasing function of q and the result follows.

b) From (A.2), differentiating with respect to ϵ, we get the following first order condition:

$$\frac{\partial E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\tau})]}{\partial\epsilon}=-(c-v)-\sigma(p-v)\int_{-\infty}^{\infty}L(z)\frac{(q-q^*-\epsilon)}{\tau^3\sqrt{2\pi}}e^{-\frac{(q-q^*-\epsilon)^2}{2\tau^2}}\,\mathrm{d}q=0.$$

Integrating by parts, we get:

$$\begin{aligned}\frac{\partial E[\Pi(\tilde{y},\tilde{q}_{\epsilon,\tau})]}{\partial\epsilon}&=-(c-v)-\sigma(p-v)\int_{-\infty}^{\infty}\frac{\partial L(z)}{\partial q}\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{(q-q^*-\epsilon)^2}{2\tau^2}}\,\mathrm{d}q\\&=-(c-v)+(p-v)-(p-v)\int_{-\infty}^{\infty}\Phi(z)\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{(q-q^*-\epsilon)^2}{2\tau^2}}\,\mathrm{d}q\\&=-(c-v)+(p-v)-(p-v)P(Y\le 0)\\&=0.\end{aligned}$$

Equivalently, $P(Y\le 0)=CR$, where Y is a normally distributed random variable with mean $\frac{\mu-q^*-\epsilon}{\sigma}$ and standard deviation $\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}$. Then

$$0=\frac{\mu-q^*-\epsilon^*}{\sigma}+z_{CR}\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}\;\Leftrightarrow\;\epsilon^*=z_{CR}\left(\sqrt{\sigma^2+\tau^2}-\sigma\right).$$

We used the fact that $q^*=\mu+z_{CR}\sigma$. Note also that $\frac{\partial^2 E[\Pi(\tilde{y},q(\epsilon))]}{\partial\epsilon^2}\le 0$, so ϵ^* maximizes (A.2). ■

PROOF OF COROLLARY 2:

Proof: Conditional on μ, $\overline{x}\sim N(\mu,\sigma^2/k^*)$. And, hence, conditional on μ, each of the three order quantities are normally distributed with variance σ^2/k^*. Moreover, $E[\tilde{q}_{CE}|\mu]=q^*$, $E[\tilde{q}_{PD}|\mu]=q^*+z_{CR}\sigma\left(\sqrt{1+\frac{1}{k^*}}-1\right)$ and $E[\tilde{q}_{PD_0}|\mu]=q^*+z_{CR}\sigma\left(\sqrt{1+\frac{1}{k}}-1\right)$. Also, $0\le|\epsilon|=\left|z_{CR}\sigma\left(\sqrt{1+\frac{1}{k}}-1\right)\right|\le|\epsilon^*|=\left|z_{CR}\sigma\left(\sqrt{1+\frac{1}{k^*}}-1\right)\right|$, because $k\ge k^*$. Then, from Theorem 1, it follows that $E[\Pi(\tilde{y},\tilde{q}_{PD})|\mu]\ge E[\Pi(\tilde{y},\tilde{q}_{PD_0})|\mu]\ge E[\Pi(\tilde{y},\tilde{q}_{CE})|\mu]$. Hence, for any prior density function on μ, the result follows. ■

B. Normal Model with Heterogeneous ρ_{ij}s

We extend §2.1 by allowing heterogeneous ρ_{ij}s in $\mathbf{\Sigma}$. In this case, the covariance matrix $\mathbf{\Sigma}$ has diagonal elements σ^2 and off-diagonal elements $\rho_{ij}\sigma^2$, with $\rho_{ij} = Corr[x_i,x_j|\theta]$, $i \neq j$. Assume ρ_{ij}s are known.

Setting $\mathbf{\Sigma} = \sigma^2 \mathbf{\Sigma}_{\rho_{ij}}$ where $\mathbf{\Sigma}_{\rho_{ij}}$ is a $k \times k$ is a matrix with diagonal elements 1 and off-diagonal elements ρ_{ij}. With $\lambda = 1/\sigma^2$, the likelihood function can be rewritten as

$$l(\boldsymbol{x}|\mu,\lambda) \propto \lambda^{\frac{k}{2}} exp\left(-\frac{\lambda k^*}{2}(\mu-\hat{\mu})^2\right) exp\left(-\frac{\lambda}{2}(k-1)\hat{s}^2\right), \tag{B.1}$$

where $k^* = \boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\boldsymbol{e}$, $\hat{\mu} = \frac{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\boldsymbol{x}}{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\boldsymbol{e}}$ and $\hat{s}^2 = \frac{(\boldsymbol{x}-\hat{\mu}\boldsymbol{e})'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}(\boldsymbol{x}-\hat{\mu}\boldsymbol{e})}{k-1}$.

Using the normal-gamma prior distribution on μ and λ in §2.1, the posterior distribution of μ and λ is given by

$$f(\mu,\lambda|\boldsymbol{x}) = NG(\mu,\lambda|\mu^*,n_\mu^*,v^*,n_v^*) = N(\mu|\mu^*,(n_\mu^*\lambda)^{-1})Ga(\lambda|\frac{n_v^*}{2},\frac{n_v^* v^*}{2}), \tag{B.2}$$

where $n_\mu^* = n_\mu + k^*$, $\mu^* = \frac{n_\mu\mu_0 + k^*\hat{\mu}}{n_\mu^*}$, $n_v^* = n_v + k$ and $v^* = \frac{1}{n_v^*}\left(n_v v_0 + (k-1)\hat{s}^2 + \frac{n_\mu k^*}{n_\mu + k^*}(\hat{\mu}-\mu_0)^2\right)$.

It then follows that $f(\tilde{y}|\boldsymbol{x})$ is a t distribution with degrees of freedom n_v^*, location parameter μ^*, and scale parameter $\sqrt{(n_\mu^*+1)v^*/n_\mu^*}$, so that $E(\tilde{y}|\boldsymbol{x}) = \mu^*$ for $n_v^* > 1$ and, for $n_v^* > 2$, $Var(\tilde{y}|\boldsymbol{x}) = \frac{n_v^*}{n_v^*-2}\left(v^* + \frac{v^*}{n_\mu^*}\right)$.

With a diffuse prior on μ and σ^2 (i.e., with $n_\mu = n_v = 0$), $f(\tilde{y}|\boldsymbol{x})$ is a t distribution with k degrees of freedom, location $\hat{\mu}$, and scale $\sqrt{(k^*+1)v^*/k^*}$, which yields $E(\tilde{y}|\boldsymbol{x}) = \hat{\mu}$ and

$$Var(\tilde{y}|\boldsymbol{x}) = \left(1+\frac{1}{k^*}\right)\frac{(\boldsymbol{x}-\hat{\mu}\boldsymbol{e})'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}(\boldsymbol{x}-\hat{\mu}\boldsymbol{e})}{k-2} \tag{B.3}$$

C. Normal Model with Heterogeneous ρ_{ij}s: Impact of μ and σ^2 on δ

Consider $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{z}\sigma$, where $\boldsymbol{z} = (z_1,...,z_k)'$ follows a multivariate normal with mean vector 0 and a covariance matrix $\mathbf{\Sigma}_{\rho_{ij}}$ as in Appendix *B*. Substituting $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{z}\sigma$ into (B.3), we obtain

$$\begin{aligned} Var(\tilde{y}|\boldsymbol{x}) &= \left(1+\frac{1}{k^*}\right)\frac{\left(\boldsymbol{\mu}+\boldsymbol{z}\sigma - \frac{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}(\boldsymbol{\mu}+\boldsymbol{z}\sigma)}{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\boldsymbol{e}}\right)'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\left(\boldsymbol{\mu}+\boldsymbol{z}\sigma - \frac{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}(\boldsymbol{\mu}+\boldsymbol{z}\sigma)}{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\boldsymbol{e}}\right)}{k-2} && \text{(C.1)} \\ &= \left(1+\frac{1}{k^*}\right)\frac{(\boldsymbol{z}-\hat{\mu}_z\boldsymbol{e})'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}(\boldsymbol{z}-\hat{\mu}_z\boldsymbol{e})}{k-2}\sigma^2, && \text{(C.2)} \end{aligned}$$

where $\hat{\mu}_z = \frac{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\boldsymbol{z}}{\boldsymbol{e}'\mathbf{\Sigma}_{\rho_{ij}}{}^{-1}\boldsymbol{e}}$. Similarly, we can apply the same transformation $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{z}\sigma$ to sample variance $s^2 = \frac{\sum_{i=1}^k (x_i-\bar{x})^2}{k-1} = \frac{\sum_{i=1}^k (\mu+z_i\sigma-(\mu+\bar{z}\sigma))^2}{k-1} = \frac{\sum_{i=1}^k (z_i-\bar{z})^2}{k-1}\sigma^2$, where $\bar{z} = \frac{1}{k}\sum_{i=1}^k z_i$. It is evident that the ratio between $Var(\tilde{y}|\boldsymbol{x})$ and s^2 is independent of μ and σ^2.

D. Lognormal Model: Impact of μ on Rescaled Interval Width

In the lognormal model in §4.2, consider $\boldsymbol{x_t} = \boldsymbol{\mu} + \boldsymbol{z_t}\sigma$, where $\boldsymbol{z_t} = (\ln z_1,...,\ln z_k)$ follows a multivariate normal with mean vector 0 and a covariance matrix with diagonal elements σ^2 and off-diagonal elements $\rho_L\sigma^2$.

By substituting $\overline{x}_t = \mu + \overline{z}_t\sigma$, where $\overline{z}_t = \frac{1}{k}\sum_{i=1}^{k} \ln z_i$, the widths of prediction intervals under PD and CE_0 (Table 7) are respectively given by:.

$$exp\left(\mu + \overline{z}_t\sigma \pm t_{1-\frac{\gamma}{2},k}\sqrt{\frac{k-1}{k}(\frac{1+\rho_L}{1-\rho_L}+\frac{1}{k})s_t}\right), \text{and} \tag{D.1}$$

$$exp\left(\mu + \overline{z}_t\sigma \pm z_{1-\frac{\gamma}{2}}s_t\right). \tag{D.2}$$

Since s_t is also independent of μ, the ratio between (D.1) and (D.2) is independent of μ. Similar argument holds for the ratio of prediction intervals widths of PD_0 and CE_0, and of CE and CE_0.

Acknowledgments

The authors are grateful to Steve Chick, Kevin McCardle, Serguei Netessine, and Bob Winkler for helpful suggestions. The Centre for Decision Making and Risk Analysis (CDMRA) provided support for this project.

References

Ashton, R. H. 1986. Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes* **38**(3) 405–414.

Bunn, D. W. 1985. Statistical efficiency in the linear combination of forecasts. *International Journal of Forecasting* **1**(2) 151–163.

Chhibber, S., G. Apostolakis. 1993. Some approximations useful to the use of dependent information sources. *Reliability Engineering & System Safety* **42**(1) 67–86.

Choi, T., D. Li, H. Yan. 2008. Mean-variance analysis for the newsvendor problem. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **38**(5) 1169–1180.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting* **5**(4) 559–583.

Clemen, R. T., G. W. Fischer, R. L. Winkler. 2000. Assessing dependence: Some experimental results. *Management Sci.* **46**(8) 1100–1115.

Clemen, R. T., R. L. Winkler. 1985. Limits for the precision and value of information from dependent sources. *Oper. Res.* **33**(2) 427–442.

Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* **44**(1) 87–99.

Fisher, M., A. Raman. 2010. *The New Science of Retailing: How Analytics are Transforming the Supply Chain and Improving Performance*. Harvard Business Review Press.

Fujikoshi, Y., S. Mukaihata. 1993. Approximations for the quantiles of student's t and f distributions and their error bounds. *Hiroshima Math. J* **23**(3) 557–564.

Gaur, V., S. Kesavan, A. Raman, M. Fisher. 2007. Estimating demand uncertainty using judgmental forecasts. *Manufacturing Service Oper. Management* **9**(4) 480–491.

Geisser, S. 1965. A bayes approach for combining correlated estimates. *Journal of the American Statistical Association* **60**(310) 602–607.

Gokhale, D. V., P. S. James. 1982. Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *Journal of the Royal Statistical Society. Series A (General)* 237–249.

Hammond, J. H., A. Raman. 1994. *Sport Obermeyer, Ltd.* HBS Case 695-022, Harvard Business School, Boston.

Johnson, N. L., S. Kotz, N. Balakrishnan. 2002. *Continuous Multivariate Distributions, Volume 1, Models and Applications*. New York: John Wiley & Sons.

Kremer, M., E. Siemsen, D. J. Thomas. 2015. The sum and its parts: Judgmental hierarchical forecasting. *Management Sci.* .

Lichtendahl Jr, K. C., Y. Grushka-Cockayne, P. E. Pfeifer. 2013. The wisdom of competitive crowds. *Oper. Res.* **61**(6) 1383–1398.

Meyer, M. A., J. M. Booker. 2001. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. SIAM.

Özer, Ö., Y. Zheng, K. Chen. 2011. Trust in forecast information sharing. *Management Sci.* **57**(6) 1111–1137.

Rao, C. R. 1973. *Linear Statistical Inference and its Applications*. John Wiley & Sons.

Schmittlein, D. C., J. Kim, D. G. Morrison. 1990. Combining forecasts: Operational adjustments to theoretically optimal rules. *Management Sci.* **36**(9) 1044–1056.

Schweitzer, M. E., G. P. Cachon. 2000. Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Sci.* **46**(3) 404–420.

Seifert, M., E. Siemsen, A. L. Hadida, A. B. Eisingerich. 2015. Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management* **36** 33–45.

Shumsky, R. A. 1998. Optimal updating of forecasts for the timing of future events. *Management Sci.* **44**(3) 321–335.

Winkler, R. L. 1981. Combining probability distributions from dependent information sources. *Management Sci.* **27**(4) 479–488.

Winkler, R. L., R. T. Clemen. 1992. Sensitivity of weights in combining forecasts. *Oper. Res.* **40**(3) 609–614.