

Forecasting and Dynamic Updating of Uncertain Arrival Rates to A Call Center

Haipeng Shen

Dept. of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA

Jianhua Z. Huang

Dept. of Statistics
Texas A&M University
College Station, TX 77843 USA

Chihoon Lee

Dept. of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA

Abstract—Motivated by queueing models and recent empirical studies of call centers, we model call arrival processes as inhomogeneous Poisson processes. Our primary interest lies on forecasting the *unobserved* intraday call rate profile using the historical call volume data. We develop methods for both interday forecasting and dynamic intraday updating of call arrival rates. Such forecasts are of great importance for effective call center workforce management. Our methods combine the data-driven approach in Shen and Huang (2007) [9] with the model-driven approach in Weinberg et al. (2007) [10]. A Poisson factor model is first formulated to achieve dimension reduction. We then describe how the estimated model can be used to provide interday forecasting as well as intraday updating. Our methods show very promising results in an application to real call center data.

I. INTRODUCTION

Call centers provide an increasingly popular example of modern service networks where customers contact their service providers through telephones and wait in *tele* queues [5], [2], [1]. According to [6], every FORTUNE 500 company has at least one call center, which employs on average 4,500 agents. Call center agents represent 4% of the workforce in US, 1.8% in Israel and 1.5% in UK. More than \$300 billion is spent annually on call centers worldwide, and about 70% of the money is human resource cost, such as agent hiring/training expense, compensation and benefit.

Hence efficient agent staffing and scheduling is one really crucial operational problem, which follows from accurate prediction of future call arrival rates and volumes. Common practice uses queueing models, and models call arrival processes using inhomogeneous Poisson processes. (See [5], [1] for tutorials and extensive review of telephone call centers research.) From queueing theory perspective, the primary interest, for call center management, is to forecast the underlying interday/intraday call *rate* profile (rather than call *volume* profile), using the historical call volume data at hand.

Usually two kinds of forecasts are needed by call center managers: 1. forecast the call rates days or weeks ahead; 2. on a particular day, dynamically update the forecast using newly available information as additional calls arrive throughout the day. Various studies such as [2], [10], [9] suggest that the underlying call arrival rate profiles exhibit significant interday

time series correlation as well as intraday time-varying trend. Such a two-way variation structure needs to be taken into account in order for a forecasting model to be successful.

In the literature, only two papers [10], [9] have considered both interday forecasting and intraday dynamic updating. Bayesian forecasting *model-driven* methods are proposed in [10] under the assumption that the call volumes follow inhomogeneous Poisson processes. For parameter estimation and forecasting, a two-way multiplicative Bayesian Gaussian model is suggested and a Markov Chain Monte Carlo (MCMC) algorithm is developed. Implementation of the MCMC algorithm is sophisticated, and the algorithm sometimes can require a long time for convergence. On the other hand, data-driven methods are developed in [9] for interday and dynamic intraday forecasting of future call volumes (not call *rates*). Treating the intraday call volume profiles as a high dimensional vector time series, the approach is to first reduce the dimensionality by singular value decomposition (SVD) of the matrix of historical intraday profiles and then apply time series regression. The *data-driven* nature means that the assumption of Poisson processes is not necessary.

In this paper we aim at extending the methods in [9] by providing forecasts for future underlying call *rate* profiles. Similar to [2], [10], we assume the call volumes are realizations of inhomogeneous Poisson processes. Different from them, we model the call volumes directly instead of the square-root-transformed volumes. The key idea of [9] is dimension reduction through SVD on the historical count matrix. We first extend SVD to the context of Poisson random variables. A low-dimensional factor models for Poisson observations is considered. The model is then estimated through an alternating sequence of Poisson regression in the context of generalized linear models (GLM) [3]. Once the model is fitted, we then propose to forecast the factor loadings, which in turn provides the forecast for future interday rate profiles. The procedure for intraday dynamic updating is based on a *penalized maximum likelihood* technique.

The proposed dimension reduction technique can be viewed as an extension of SVD for Poisson data. The alternating fitting algorithm is very easy to implement. The forecasting approach enjoys the benefit of both the model-

driven approach of [10] and the data-driven approach of [9]. It directly models and forecasts the rate profiles. Our methods generate promising results in a real application.

The rest of the paper is organized as follows. Section II introduces the call center arrival data. Our dimension reduction approach is described in detail in Section III. Section III-A reviews the basic ideas of SVD and Section III-B introduces the Poisson factor model. Section III-C then provides an alternating estimation algorithm. Section IV describes forecasting methods for one- or multi-day-ahead intraday call rate profile forecasting (Section IV-A) as well as dynamic intraday rate updating (Section IV-B). Section V uses the data described in Section II to illustrate our approach and compare it with existing methods. We conclude the paper in Section VI with a discussion of future work.

II. THE DATA

The motivating data are the same one studied in [9]. They were gathered at an inbound call center of a major northeastern US financial firm between January 1 and October 26, 2003. The center is normally open between 7:00AM and midnight. We focus on the 42 whole 5-day weeks between January 6 and October 24. After excluding obvious abnormal days in the data, e.g. six holidays with very low call volumes and four days with no data, and dividing the 17-hour operating period into sixty-eight 15-minute interval, the aggregated data form a 200×68 count matrix: Each row corresponding to a particular day in the 42 weeks considered, and each column corresponding to one specific 15-minute interval between 7:00AM and midnight. (See Section 2 in [9] for more details/features of this data.)

III. THE MODEL

Let $\mathbf{Y} = (y_{ij})$ be an $n \times m$ matrix that records the call volumes for n days with each day having m time periods. The rows and columns of \mathbf{Y} correspond respectively to days and time periods within a day. The i th row of \mathbf{Y} , denoted as $\mathbf{y}_{(i)}^T = (y_{i1}, \dots, y_{im})$, is referred to as the *intraday call volume profile* of the i th day. The j th column of \mathbf{Y} , denoted as $\mathbf{y}_j = (y_{1j}, \dots, y_{nj})^T$, is referred to as the *interday call volume profile* of the j th time period. Furthermore, we assume that y_{ij} is a Poisson random variable with rate λ_{ij} . For notational purpose, we let $\Lambda = (\lambda_{ij})$ denote the $n \times m$ matrix containing the set of Poisson rates. The i th row of Λ , denoted as $\boldsymbol{\lambda}_{(i)}^T = (\lambda_{i1}, \dots, \lambda_{im})$, is referred to as the *intraday call rate profile* of the i th day. The j th column of Λ , denoted as $\boldsymbol{\lambda}_j = (\lambda_{1j}, \dots, \lambda_{nj})^T$, is the *interday call rate profile* of the j th time period.

The intraday rate profiles, $\{\boldsymbol{\lambda}_{(1)}, \boldsymbol{\lambda}_{(2)}, \dots\}$, form a vector-valued time series taking values in \mathbb{R}^m . We want to build a time series model for this series and use it for forecasting. However, the rate profiles are not directly observable, which makes the forecasting problem difficult. The idea is to build a forecasting model on the count profiles, $\{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \dots\}$, and use it to forecast future rate profiles.

In practice the dimensionality of the count profiles is usually so large that it is infeasible to directly apply commonly used multivariate time series models such as VAR (Vector Autoregressive models) and more general VARMA (Vector Autoregressive and Moving Average models) [7]. For example, the dimensionality m is 68 for our data and 169 for the application in [10]. This calls for the necessity of dimension reduction. In addition, each $\mathbf{y}_{(i)}$ is a Poisson random vector with a *positive* rate vector $\boldsymbol{\lambda}_{(i)}$. This distributional nature needs to be accounted for appropriately.

A. Dimension Reduction via Singular Value Decomposition

To achieve dimension reduction, Shen and Huang (2007) [9] consider the following decomposition of the square-root-transformed count matrix $\mathbf{X} = \sqrt{\mathbf{Y} + 1/4}$,

$$\mathbf{x}_{(i)} = \beta_{i1} \mathbf{f}_1 + \dots + \beta_{iK} \mathbf{f}_K + \boldsymbol{\epsilon}_{(i)}, \quad i = 1, \dots, n, \quad (1)$$

where $\mathbf{f}_1, \dots, \mathbf{f}_K \in \mathbb{R}^m$ are the basis vectors and $\boldsymbol{\epsilon}_{(1)}, \dots, \boldsymbol{\epsilon}_{(n)} \in \mathbb{R}^m$ are the error terms. For a fixed K , they expect that $\mathbf{x}_{(i)}$ can be well summarized by the linear combination of the basis vectors so that the error terms in (1) would be small in magnitude. Thus, the model can be estimated by solving the following minimization problem,

$$\min_{\substack{\beta_{i1}, \dots, \beta_{iK} \\ \mathbf{f}_1, \dots, \mathbf{f}_K}} \sum_{i=1}^n \|\boldsymbol{\epsilon}_{(i)}\|^2 \equiv \sum_{i=1}^n \left\| \mathbf{x}_{(i)} - (\beta_{i1} \mathbf{f}_1 + \dots + \beta_{iK} \mathbf{f}_K) \right\|^2. \quad (2)$$

For identifiability, it is required that $\mathbf{f}_i^T \mathbf{f}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta which equals 1 for $i = j$ and 0 otherwise. The solution to this problem is actually given by the SVD of the matrix \mathbf{X} as shown below.

The SVD of the matrix \mathbf{X} can be expressed as

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T, \quad (3)$$

where \mathbf{U} is an $n \times m$ matrix with orthonormal columns, \mathbf{S} is an $m \times m$ diagonal matrix, and \mathbf{V} is an $m \times m$ orthogonal matrix. The triple $(\mathbf{U}, \mathbf{V}, \mathbf{S})$ contains the SVD components of \mathbf{X} . The columns of \mathbf{U} , $\{\mathbf{u}_k = (u_{1k}, \dots, u_{nk})^T\}$, namely the left singular vectors, satisfy $\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$. The columns of \mathbf{V} , $\{\mathbf{v}_k = (v_{1k}, \dots, v_{mk})^T\}$, or the right singular vectors, satisfy $\mathbf{v}_i^T \mathbf{v}_j = \delta_{ij}$. The diagonal elements of \mathbf{S} are the singular values, which are usually ordered decreasingly. Let $\mathbf{S} = \text{diag}(s_1, \dots, s_m)$ and $r = \text{rank}(\mathbf{X})$. Then $s_1 \geq s_2 \geq \dots \geq s_r > 0$, and $s_k = 0$ for $r + 1 \leq k \leq m$.

It then follows from (3) that

$$\mathbf{x}_{(i)} = s_1 u_{i1} \mathbf{v}_1 + \dots + s_r u_{ir} \mathbf{v}_r.$$

Keeping only the terms associated with the largest K singular values, we have the following approximation,

$$\mathbf{x}_{(i)} \simeq (s_1 u_{i1}) \mathbf{v}_1 + \dots + (s_K u_{iK}) \mathbf{v}_K.$$

This K -term approximation is an optimal solution for the minimization problem (2) [4]. More precisely, $\beta_{ik} = s_k u_{ik}$ and $\mathbf{f}_k = \mathbf{v}_k$, $i = 1, \dots, n$, $k = 1, \dots, K$, solve (2).

The authors illustrate using case studies that $K = 2$ or 3 is sufficient to achieve very good forecasting performance. A substantial dimension reduction is obtained considering the large original dimensionality ($m = 68$ or 169).

B. Poisson Singular Value Decomposition

The square-root transformation in [9] aims at stabilizing the heteroscedastic variances. The same transformation has been utilized in [2] and [10]. They make the Poisson assumption, under which the square-root transformation not only stabilizes the variance, but also makes the transformed counts approximately normally distributed.

Instead of the square-root-transformed count matrix \mathbf{X} , we would like to model the original count matrix \mathbf{Y} and the rate matrix Λ . We start by dimension reduction of the Poisson rate matrix Λ , or some transformation of it such as $\sqrt{\Lambda}$. Here the notation $\sqrt{\Lambda}$ represents a matrix whose entries are the square-root transformation of the corresponding entries of the matrix Λ . Let g denote a general, fixed transformation function, called the link function in the generalized linear model (GLM) literature [3]. Denote by $g(\lambda_{(i)})$ the vector $(g(\lambda_{i1}), \dots, g(\lambda_{im}))^T$.

Extending SVD, we would like to seek a few underlying factors in \mathbb{R}^m , denoted as \mathbf{f}_k , $k = 1, \dots, K$, such that all elements in the time series $\{g(\lambda_{(i)})\}$ can be represented by these factors. The number of factors K should be much smaller than the dimensionality m of the time series. Specifically, we consider the following factor model,

$$g(\lambda_{(i)}) = \beta_{i1}\mathbf{f}_1 + \dots + \beta_{iK}\mathbf{f}_K = \mathbf{F}\boldsymbol{\beta}_{(i)}, \quad i = 1, \dots, n, \quad (4)$$

where $\mathbf{F}_{m \times K} = (\mathbf{f}_1, \dots, \mathbf{f}_K)$ is the factor matrix and $\boldsymbol{\beta}_{(i)} = (\beta_{i1}, \dots, \beta_{iK})^T$ is the factor loading vector for the i th intraday rate profile. Denote $\mathbf{B}_{n \times K} = (\boldsymbol{\beta}_{(1)}, \dots, \boldsymbol{\beta}_{(n)})^T$. Then the model (4) can be written in matrix form as

$$g(\Lambda) = \mathbf{B}\mathbf{F}^T.$$

For later use, we let β_1, \dots, β_K denote the columns of \mathbf{B} . For identifiability, we also require in (4) that $\mathbf{f}_k^T \mathbf{f}_{k'} = \delta_{kk'}$. Note that requiring $\mathbf{f}_k^T \mathbf{f}_{k'} = \delta_{kk'}$ is only one possible way to make the model identifiable. Alternatively, we can require that $\sum_{i=1}^n \beta_{ik}\beta_{ik'} = \delta_{kk'}$. This orthogonality requirement has important implications in constructing our forecasting models as described below in Section IV-A.

One attractive character of the model (4) is that it effectively separates out the intraday and interday variations, both of which are present in the intraday rate profile time series. Our proposal is to first extract the intraday factors and their loadings using historical data by fitting the factor model (4). (See Section III-C for an iterative fitting algorithm.) Secondly, time series models are built on the loading time series to derive forecasts for future loadings, which can then be combined with the factors to forecast future rate profiles (Section IV-A). Finally, these forecasts can be combined with the intraday factors to update latter intraday profiles using newly observed limited count profiles (Section IV-B).

C. An Alternating Estimation Algorithm

If the Poisson rate profiles $\lambda_{(i)}$'s were observable, the factors and their loadings can be easily extracted using singular value decomposition [8], [9]. Because they are unobserved, in this section, we derive an iterative procedure to extract the factors/loadings. Our procedure is based on the observation that when the factors are fixed, the model (4) implies Poisson regression models for the loading vectors and vice versa.

Assuming the factor model (4) holds, we have the following Poisson regression model for the i th intraday call volume profile $\mathbf{y}_{(i)}$,

$$\begin{cases} \mathbf{y}_{(i)} \sim \text{Poisson}(\lambda_{(i)}), \\ g(\lambda_{(i)}) = \mathbf{F}\boldsymbol{\beta}_{(i)}, \quad i = 1, \dots, n, \end{cases} \quad (5)$$

where $\mathbf{y}_{(i)} \sim \text{Poisson}(\lambda_{(i)})$ means that $y_{ij} \sim \text{Poisson}(\lambda_{ij})$ for $1 \leq j \leq m$, and $g(\lambda_{(i)}) = (g(\lambda_{i1}), \dots, g(\lambda_{im}))^T$.

Similarly, the j th interday call volume profile \mathbf{y}_j satisfies the Poisson regression model below,

$$\begin{cases} \mathbf{y}_j \sim \text{Poisson}(\lambda_j), \\ g(\lambda_j) = \mathbf{B}\mathbf{f}_{(j)}, \quad j = 1, \dots, m, \end{cases} \quad (6)$$

where $\mathbf{y}_j \sim \text{Poisson}(\lambda_j)$ means that $y_{ij} \sim \text{Poisson}(\lambda_{ij})$ for $1 \leq i \leq n$, and $g(\lambda_j) = (g(\lambda_{1j}), \dots, g(\lambda_{nj}))^T$. Note that $\mathbf{f}_{(j)}^T$ is the j th row of \mathbf{F} .

Both models (5) and (6) can be estimated by maximizing the corresponding likelihood functions. The above discussion suggests an alternating maximum likelihood algorithm to extract the factor matrix \mathbf{F} and the loading matrix \mathbf{B} .

IV. FORECASTING METHODS

A. Interday Forecasting

Queueing theory suggests that, for call center management, the primary interest is to forecast the intraday call rate profile $\lambda_{(n+h)}$, using the historical call volume data $\{\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(n)}\}$. The point forecast for the intraday call volume profile $\mathbf{y}_{(n+h)}$ and its rate are the same; however their distributional forecasts are different. We currently focus on the point forecast only.

As suggested by the model (4), forecasting a m -dimensional time series $\{\lambda_{(i)}\}$ reduces to forecasting K one-dimensional interday feature series β_1, \dots, β_K . We consider the factor model (4) with the factors \mathbf{f}_k 's replaced by the ones estimated from the historical data matrix \mathbf{Y} . Because of the way the model is estimated, it is adequate to forecast each β_k separately using univariate time series methods.

If we can obtain a forecast of the factor loadings vector $\boldsymbol{\beta}_{(n+h)} = \{\beta_{n+h,1}, \dots, \beta_{n+h,K}\}^T$, then a point forecast of $\lambda_{(n+h)}$ is given by

$$\hat{\lambda}_{(n+h)} = g^{-1}(\hat{\beta}_{n+h,1}\mathbf{f}_1 + \dots + \hat{\beta}_{n+h,K}\mathbf{f}_K),$$

where $\hat{\beta}_{n+h,k}$ is a forecast of $\beta_{n+h,k}$, $k = 1, \dots, K$. Because $\mathbf{y}_{(n+h)}$ has the Poisson distribution with rate $\lambda_{(n+h)}$, the point forecast of the call volume profile $\mathbf{y}_{(n+h)}$ is the same as the rate profile point forecast, $\hat{\lambda}_{(n+h)}$.

B. Dynamic Intraday Updating

Using the method in the previous section, a call center manager can forecast the rate profile for day $n+1$ at the end of day n . As calls arrive during day $n+1$, the manager may want to dynamically update her forecast for the remainder of the day using the information from the early part of that day. This dynamic updating is useful because it adds flexibility to her allocation of available resources, which then leads to higher efficiency and productivity, as well as better quality of service. By reevaluating her forecast for the remainder of the day, she can then schedule meetings or training sessions for agents free from work at short-notice, or call in back-up agents.

Suppose we have available the call volumes during the first m_0 time periods of day $n+1$. Denote them collectively as $\mathbf{y}_{(n+1)}^e = (y_{n+1,1}, \dots, y_{n+1,m_0})^T$, a vector containing the first m_0 elements of $\mathbf{y}_{(n+1)}$ with the corresponding Poisson rate vector being $\boldsymbol{\lambda}_{(n+1)}^e$. Denote $\mathbf{y}_{(n+1)}^l = (y_{n+1,m_0+1}, \dots, y_{n+1,m})^T$ to be the intraday call volume profile for the latter part of day $n+1$ with the rate vector being $\boldsymbol{\lambda}_{(n+1)}^l$. For notational simplicity, we suppress the dependence of $\mathbf{y}_{(n+1)}^e/\boldsymbol{\lambda}_{(n+1)}^e$ and $\mathbf{y}_{(n+1)}^l/\boldsymbol{\lambda}_{(n+1)}^l$ on m_0 .

Let $\hat{\beta}_{n+1,k}^{\text{TS}}$ be a time series (TS) forecast based on β_k , for $k = 1, \dots, K$, using information up to the end of day n . The TS point forecast of $\mathbf{y}_{(n+1)}^l$ and $\boldsymbol{\lambda}_{(n+1)}^l$ is then given by for $j = m_0 + 1, \dots, m$,

$$\hat{y}_{n+1,j}^{\text{TS}} = \hat{\lambda}_{n+1,j}^{\text{TS}} = g^{-1}(\hat{\beta}_{n+1,1}^{\text{TS}} f_{j1} + \dots + \hat{\beta}_{n+1,K}^{\text{TS}} f_{jK}).$$

These forecasts do not utilize any new information of day $n+1$. Below, we discuss two ways to incorporate the new information in $\mathbf{y}_{(n+1)}^e$ to obtain an updated forecast of $\mathbf{y}_{(n+1)}^l$ and $\boldsymbol{\lambda}_{(n+1)}^l$.

1) *Direct Updating*: When being applied to the intraday rate profile of day $n+1$, the factor model (4) can be written as

$$g(\lambda_{n+1,j}) = \beta_{n+1,1} f_{j1} + \dots + \beta_{n+1,K} f_{jK}, \quad j = 1, \dots, m.$$

Let \mathbf{F}^e be a $m_0 \times K$ matrix whose (j,k) th entry is f_{jk} , $1 \leq j \leq m_0$, $1 \leq k \leq K$, $\boldsymbol{\beta}_{(n+1)} = (\beta_{n+1,1}, \dots, \beta_{n+1,K})^T$. Then, with the availability of $\mathbf{y}_{(n+1)}^e$, we have the following Poisson regression model,

$$\begin{cases} \mathbf{y}_{(n+1)}^e \sim \text{Poisson}(\boldsymbol{\lambda}_{(n+1)}^e) \\ g(\boldsymbol{\lambda}_{(n+1)}^e) = \mathbf{F}^e \boldsymbol{\beta}_{(n+1)}. \end{cases} \quad (7)$$

This suggests that we can forecast $\boldsymbol{\beta}_{(n+1)}$ by the method of maximum likelihood (ML), solving the following optimization problem,

$$\begin{aligned} \min_{\boldsymbol{\beta}_{(n+1)}} \sum_{j=1}^{m_0} [\lambda_{n+1,j} - y_{n+1,j} \log(\lambda_{n+1,j})] \\ \text{subject to } g(\boldsymbol{\lambda}_{(n+1)}^e) = \mathbf{F}^e \boldsymbol{\beta}_{(n+1)}, \end{aligned}$$

which is equivalent to minimize over $\boldsymbol{\beta}_{(n+1)}$

$$\sum_{j=1}^{m_0} \left\{ g^{-1} \left(\sum_{k=1}^K \beta_{n+1,k} f_{jk} \right) - y_{n+1,j} \log \left[g^{-1} \left(\sum_{k=1}^K \beta_{n+1,k} f_{jk} \right) \right] \right\}.$$

Note that the above optimization criterion is actually the negative log-likelihood function of $\mathbf{y}_{(n+1)}^e$.

Using numerical algorithms such as Newton-Raphson or Fisher Scoring, we can obtain $\hat{\boldsymbol{\beta}}_{(n+1)}^{\text{ML}}$. Hence, the ML point forecast of $\mathbf{y}_{(n+1)}^l$ and $\boldsymbol{\lambda}_{(n+1)}^l$ is then given by for $j = m_0 + 1, \dots, m$,

$$\hat{y}_{n+1,j}^{\text{ML}} = \hat{\lambda}_{n+1,j}^{\text{ML}} = g^{-1}(\hat{\beta}_{n+1,1}^{\text{ML}} f_{j1} + \dots + \hat{\beta}_{n+1,K}^{\text{ML}} f_{jK}).$$

2) *Penalized Updating*: Direct updating only makes use of the additional information available at the early part of day $n+1$. It needs a sufficient amount of data (i.e. a large enough m_0) in order for $\hat{\boldsymbol{\beta}}_{(n+1)}^{\text{ML}}$ to be reliable. This might create a problem if the manager wants to update her forecast early in the morning, for example, at 8:00AM with $m_0 = 4$ or 10:00AM with $m_0 = 12$ for the data in Section II. Another disadvantage of the direct ML updating is that it does not make full use of the historical information other than the estimated intraday feature vectors. In particular, it ignores the day-to-day dependence present in the interday feature series.

We propose combining the maximum likelihood forecast with the time series forecast of $\boldsymbol{\beta}_{(n+1)}$ via the idea of penalization. Specifically, we minimize with respect to $\beta_{n+1,1}, \dots, \beta_{n+1,K}$ the following *penalized likelihood criterion*,

$$\sum_{j=1}^{m_0} [\lambda_{n+1,j} - y_{n+1,j} \log(\lambda_{n+1,j})] + \omega \sum_{k=1}^K |\beta_{n+1,k} - \hat{\beta}_{n+1,k}^{\text{TS}}|^2, \quad (8)$$

subject to

$$g(\boldsymbol{\lambda}_{(n+1)}^e) = \mathbf{F}^e \boldsymbol{\beta}_{(n+1)},$$

where $\hat{\beta}_{n+1,k}^{\text{TS}}$ is a time series forecast based on the information up to the end of day n , and $\omega > 0$ is a penalty parameter. Since only one penalty parameter is used, it makes sense for the K time series $\{\beta_{ik}\}$, $1 \leq k \leq K$, to be roughly on the same scale. This can be achieved by requiring $\sum_{i=1}^n \beta_{ik} \beta_{ik'} = \delta_{kk'}$ or $(1/n) \sum_{i=1}^n \beta_{ik} \beta_{ik'} = \delta_{kk'}$ in (4).

The criterion (8) involves two terms: the first term measures the goodness-of-fit of the model to the observed call volumes in the early part of the day, while the second term penalizes a large departure from the time series forecast. The $\boldsymbol{\beta}_{(n+1)}$ obtained as the solution to the minimization problem is a compromise between the two terms based on the size of ω , the penalty parameter. In practice, ω can be selected based on the forecasting performance on a rolling hold-out sample; see Section 4.4.2 in [9] for a detailed description of one selection procedure. Direct minimization of the penalized criterion (8) can be complicated. Instead we use an iterative algorithm based on the quadratic approximation of (8).

Minimizing this criterion gives us the *penalized maximum likelihood* (PML) forecast of $\beta_{(n+1)}$. The PML point forecast of $y_{(n+1)}^l$ and $\lambda_{(n+1)}^l$ is then given by for $j = m_0 + 1, \dots, m$,

$$\hat{y}_{n+1,j}^{\text{PML}} = \hat{\lambda}_{n+1,j}^{\text{PML}} = g^{-1}(\hat{\beta}_{n+1,1}^{\text{PML}} f_{j1} + \dots + \hat{\beta}_{n+1,K}^{\text{PML}} f_{jK}). \quad (9)$$

V. NUMERICAL RESULTS

A. One-day-ahead Forecasting

We consider $K = 1, \dots, 5$ and fit the corresponding factor model (5) to the call center data described in Section II. The square-root link function is employed. Fig. 1 plots the first four factors, which are normalized to achieve comparable scales. The first factor summarizes the average daily rate profile. The additional factors capture various contrasts among different time periods within a day. Together, they show that different weekday has a different arrival rate profile. The findings are consistent with [9], except theirs are about the count (instead of rate) profiles.

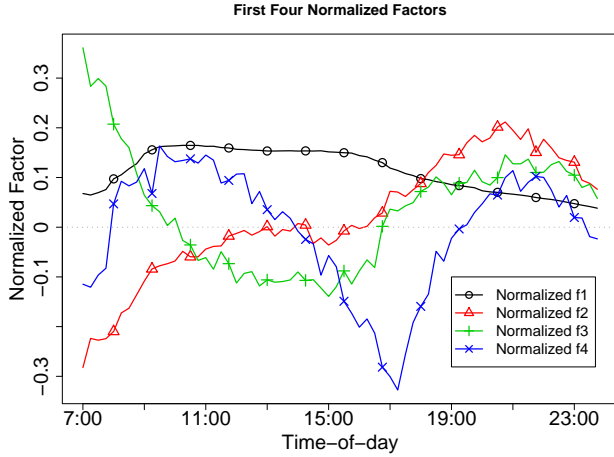


Fig. 1. The First Four Underlying Intraday Factors.

To perform interday forecasting, we need to develop some forecasting model for the loading time series. Fig. 2 plots the first loading series β_1 . Different colors and symbols indicate different weekdays, revealing a strong weekly effect and interday correlation. Further exploratory analysis motivates us to consider the following varying-coefficient AR(1) model,

$$\beta_{i1} = a_1(d_{i-1}) + b_1\beta_{i-1,1} + \eta_{i1}, \quad (10)$$

where d_{i-1} denotes the day-of-the-week of day $i - 1$, and the varying intercept a_1 depends on d_{i-1} . The same model holds for the other loading series as well.

Below we perform a rolling one-day-ahead forecasting exercise to illustrate our forecasting methods, denoted as TS1, ..., TS5 depending on the number of factor K . The last 50 days are treated as the forecasting set; for each day in the set, its preceding 150 days are used as the historical data to re-estimate the model (10) and generate the forecast.

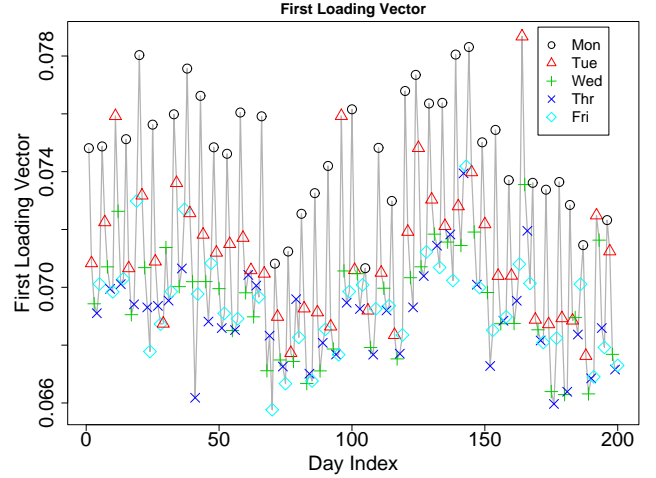


Fig. 2. Time Series Plot of the First Loading Series. This suggests an AR(1) time series model with a day-of-the-week effect.

To compare the performance of forecasting the counts, two performance measures are calculated, the root mean squared error (RMSE) and mean relative error (MRE). For day i in each data set, we define

$$\text{RMSE}_i = \sqrt{\frac{1}{m} \sum_{j=1}^m (\hat{y}_{ij} - y_{ij})^2},$$

and

$$\text{MRE}_i = \frac{100}{m} \sum_{j=1}^m \frac{|\hat{y}_{ij} - y_{ij}|}{y_{ij}},$$

where \hat{y}_{ij} is the forecast for the count y_{ij} . Forecasting errors on the underlying rates can not be calculated for real data, and we plan to investigate them using simulation studies.

Table I compares summary statistics of the RMSE and MRE of the count forecasts from the five TS methods. In general, the forecasting accuracy improves as one increases K , the number of factors. TS4 and TS5 give very comparable results. On the ground of the most parsimonious model, we decided to choose TS4.

B. Intraday Updating

Below we compare the performance of intraday updating of various methods. We look at the 10:00AM updating and the 12:00PM updating. The benchmark is the TS4 method which performs no updating. We decide to use $K = 4$ in light of the one-day-ahead forecasting performance in Section V-A. Correspondingly we term our penalized intraday updating approach as PML4.

For PML4, the value of the penalty parameter ω needs to be decided at each updating point. To this end, we consider a set of candidate values, and choose the one that minimizes some forecasting performance measure based on the call volumes (rather than rates), such as RMSE. The following

TABLE I

SUMMARY STATISTICS (MEAN, MEDIAN, LOWER QUANTILE Q1, UPPER QUANTILE Q3) OF RMSE AND MRE IN A ROLLING FORECAST EXERCISE. THE FORECASTING SET CONTAINS 50 DAYS.

	RMSE			
	Q1	Median	Mean	Q3
TS1	43.22	53.90	60.17	63.71
TS2	39.45	51.15	58.06	61.07
TS3	43.10	52.07	57.74	58.13
TS4	41.78	48.40	56.96	59.32
TS5	40.79	48.27	56.84	60.08
	MRE (%)			
	Q1	Median	Mean	Q3
TS1	5.1	6.5	7.4	8.6
TS2	4.9	5.6	6.6	7.2
TS3	4.8	5.5	6.4	6.9
TS4	4.6	5.3	6.2	6.7
TS5	4.6	5.3	6.2	6.6

out-of-sample forecast exercise is performed on the first 150 days of the data (*the training set*).

We treat the last one third (i.e., 50 days) of the training set as a rolling hold-out sample. For each day in the hold-out sample, its preceding 100 days are used to fit the Poisson factor model (5) with $K = 4$ and the square-root link function. The PML4 updating is then generated for each given ω . Let's consider, for argument's sake, the 10:00AM updating. Compute some performance measure for every day in the hold-out sample and calculate the average value. The ω that minimizes this average performance measure will be used for all days in the forecasting set. In this study, we consider ω from $\{0, 10, \dots, 10^9\}$, and $\omega = 10^3$ is chosen for both updates.

For comparison purpose, we also consider two alternative intraday updating approaches, HPM and HPA, which are studied in Section 4.2.1 of [9]. The approaches are based on historical proportions (HP), and are shown to be very competitive in [9].

Table II presents summary statistics of the RMSE of the forecasts from TS4 and PML4. The averages are calculated over the 50 days in the forecasting set. For a fair comparison, only data after 12:00PM are used when calculating the RMSE. The superior performance of PML4 is quite clear. We also observe that updating later improves the forecasting accuracy. We plan to investigate the effect of the various intraday updating methods on deciding the related staffing level.

VI. CONCLUSION

We develop a Poisson factor model that combines dimension reduction with the assumption of inhomogeneous Poisson processes. The fitted model can be used to provide interday forecasting and intraday updating of hidden rate profiles. The approach enjoys the benefits of both the data-driven approach of [9] and the model-driven approach of [10]. We illustrate the developed forecasting methods using a real

TABLE II

SUMMARY STATISTICS (MEAN, MEDIAN, LOWER QUANTILE Q1 AND UPPER QUANTILE Q3) OF RMSE FOR THE 10:00AM AND 12:00PM UPDATES. PML4 OUTPERFORMS THE OTHER METHODS.

	Q1	Median	Mean	Q3
	10:00AM updating			
TS4	36.17	41.73	52.62	55.00
HPM	36.96	45.20	55.50	61.40
HPA	37.26	49.72	56.37	62.28
PML4	36.76	43.11	50.13	52.91
12:00PM updating				
TS4	36.17	41.73	52.62	55.00
HPM	34.89	38.02	47.17	46.76
HPA	35.58	38.56	48.57	47.69
PML4	33.48	37.44	44.90	42.45

call center application, which perform well in terms of count forecasting performance.

Given the importance of forecasting rate profiles in call center operations, we plan to investigate the performance of forecasting rate profiles using simulation studies. This is necessary since we do not get to observe the underlying rates in practice. We are also interested in studying the effects of intraday updating on determining the staffing level. Another natural topic of future research is the selection of K , the number of underlying factors. The choice of the link function g is also of interest for investigation. We also plan to investigate possible ways to generate distributional forecasts and updates.

ACKNOWLEDGMENT

This work is partially supported by National Science Foundation (NSF) grant DMS-0606577 and DMS-0606580.

REFERENCES

- [1] Z. Aksin, M. Armony, and V. Mehrotra. The modern call-center: A multi-disciplinary perspective on operations management research. *unpublished*, 2007.
- [2] L. D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.
- [3] A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, 2 edition, 2001.
- [4] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.
- [5] N. Gans, G. M. Koole, and A. Mandelbaum. Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5:79–141, 2003.
- [6] P. Khudyakov. Designing a call center with an IVR. Master's thesis, Technion, 2005.
- [7] G. C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer, 2 edition, 2003.
- [8] H. Shen and J. Z. Huang. Analysis of call centre arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry*, 21:251–263, 2005.
- [9] H. Shen and J. Z. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management*, in press, 2007.
- [10] J. Weinberg, L. D. Brown, and J. R. Stroud. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, in press, 2007.