THE BEST LINEAR UNBIASED ESTIMATOR FOR CONTINUATION OF A FUNCTION

By Yair Goldberg[*], Ya'acov Ritov[*] and Avishai Mandelbaum[†]

The Hebrew University[*] and Technion-Israel Institute of Technology[†]

We show how to construct the best linear unbiased predictor (BLUP) for the continuation of a curve in a spline-function model. We assume that the entire curve is drawn from some smooth random process and that the curve is given up to some cut point. We demonstrate how to compute the BLUP efficiently. Confidence bands for the BLUP are discussed. Finally, we apply the proposed BLUP to real-world call center data. Specifically, we forecast the continuation of both the call arrival counts and the workload process at the call center of a commercial bank.

1. Introduction. Many data sets consist of a finite number of multi-dimensional observations, where each of these observations is sampled from some underlying smoothed curve. In such cases it can be advantageous to address the observations as functional data rather than as multiple series of data points. This approach was found useful, for example, in noise reduction, missing data handling, and in producing robust estimations (see the books Ramsay and Silverman, 2002, 2005, for a comprehensive treatment of functional data analysis). In this work we consider the problem of forecasting the continuation of a curve using functional data techniques.

The problem we consider here is relevant to longitudinal data sets, in which each observation consists of a series of measurements over time that describe an underlying curve. Examples of such curves are growth curves of different individuals and arrival rates of calls to a call center or of patients to an emergency room during different days. We assume that such curves, or measurement series that approximate these curves, were collected previously. We would like to estimate the continuation of a new curve given its beginning, using the behavior of the previously collected curves.

Although each observation consists of a finite number of points, the observation can be thought of as a smooth function. This dual representation leads to two different approaches when attempting to solve the prediction problem. In the discrete approach, each observation is a longitudinal vector of length $p + q$. We are interested in the prediction of the last q-length part

---

of the new observation, given its beginning p-length part. This can be computed by treating the beginning p-length vector as the predictor variables and the last q-length vector as the response variables. A prediction can be found, for example, by finding the best linear unbiased predictor (see (5)). The disadvantage of the discrete approach is that the smooth nature of the underlying function is ignored. If, instead, the continuous approach is used, the prediction problem might be treated naively using regression techniques in which both the predictor and the response are functions (Ramsay and Silverman, 2005, Chapter 16). However, these techniques do not take into account the fact that the response function is a smooth continuation of the predictor function.

In this paper, we choose the continuous approach. Specifically, we would like to generalize the discrete case to the continuous one, taking the smooth nature of the curves into account. There are three main points that need to be addressed. First, the curves lie within an infinite dimensional space, while the number of observed curves is finite. This indicates that a simple model for description of the data is required. Second, the full-length curves, the curve beginnings, and the curve continuations all lie in different functional spaces, which, in contrast to the discrete case, cannot generally be related by a linear projection. Third, we require that the prediction should be a smooth continuation of the beginning of the curve (at least in the absence of noise).

Forecasting of the continuation of a function was considered in previous works. Besse, Cardot and Stephenson (2000), and Antoniadis, Paparoditis and Sapatinas (2006), among others, developed different techniques for curve-valued autoregressive processes. In these models, each curve describes some longitudinal data cycle such as climate variation during a year (Besse, Cardot and Stephenson, 2000) and television audience rates during a day (Antoniadis, Paparoditis and Sapatinas, 2006). These models assume that the cycles behave according to some autoregressive model. The aim of these works is to predict the next cycles given past observed cycles. The continuity point at the beginning of each cycle, if it exists, is usually not taken into account. The model discussed in Shen (2009) is closer to ours. Shen discusses a curved-valued time series model in which past curves were previously observed, and the beginning part of a new curve is given. Shen first forecasts the new curve entirely, and then updates this forecast based on the given curve beginning using penalized least squares. However, all the models discussed above assume some time series behavior, while the model discussed here assumes that the curve-valued observations are independent.

The forecasting of curve continuation suggested here is based on finding

the best linear unbiased predictor (BLUP) (Robinson, 1991). We assume that the curves are governed by a small number of factors, possibly with additional noise. These factors determine the main variation between the different curves. The computation of the predictor is performed in two steps. First, the factors' coefficients are estimated from the beginning of the new curve, which is defined on the first part of the segment. Second, the prediction is obtained by computing the representation of the factors on the latter part of the segment. We prove that the resulting estimator is indeed the BLUP and that it is a smooth continuation of the beginning of the curve (at least in the absence of noise).

The two-step procedure for obtaining the BLUP involves computation of the mean function on both partial segments, and of the covariance operator on both segments and between them, which can be computationally demanding if not performed prudently. We approximate the curve data using a spline function space of (possibly large) finite-dimension (de Boor, 2001). More specifically, we represent the curves using appropriate B-splines bases. The use of splines is common in functional data analysis due to the simplicity of spline computation, and the ability of splines to approximate smooth functions. We take advantage of two more attributes of finite-dimensional spline functional spaces. First, the functional space restriction from the whole segment to a partial segment (the beginning part or the latter part) has a natural B-spline basis that has a lower number of elements. This solves collinearity problems which can render any projection on the partial segment basis instable. Second, the knot-insertion algorithm (see de Boor, 2001, Chapter 11) ensures an efficient and stable way to compute the mean function and covariance operators on different partial segments.

The proposed forecasting procedure yields a smooth curve which is the best linear unbiased prediction. Note, however, that the continuation part of the function is random, and therefore requires confidence bands. We present confidence bands for the prediction, following Knafl, Sacks and Ylvisaker (1985), under the assumption that the curves aries from a Gaussian process. The bands are computed in two steps. First, confidence intervals are computed simultaneously for a finite set of points. Then, using the fact that splines are piecewise polynomials, a global band is found. We also suggest a way to compute confidence bands using cross-validation. While no theoretical justification proof is given for the cross validation confidence bands, they are much faster to compute, and the numerical examples in Section 5 show that this approach works considerably well.

We apply the forecasting procedure suggested here to call center data. We forecast the continuation of two processes: the arrival process and the

workload process (i.e., the amount of work in the system; see, for example, Aldor-Noiman, Feigin and Mandelbaum, 2009). In call centers, the forecast of the arrival process plays an important roll in determining staffing levels. Optimization of the latter is important since salaries account for about 60-70% of the cost of running a call center (Gans, Koole and Mandelbaum, 2003). Usually, call center managers utilize forecasts of the arrival process and knowledge of service times, along with some understanding of customer patience characteristics (Zeltyn, 2005), to estimate future workload and determine staffing level (Aldor-Noiman, Feigin and Mandelbaum, 2009). The disadvantage of this approach is that the forecast of the workload is not performed directly, and instead it is obtained using the forecast of the arrival process. Reich (2010) showed how the workload process can be estimated explicitly, thereby enabling direct forecast of the workload. In this work we forecast the continuation of both the arrival and workload processes, given past days' information and the information up to some time of the day. We compare between the results for the arrival process and the workload process. We also compare our results for the arrival process to those of other forecasting techniques, namely, to the techniques that were introduced by Weinberg, Brown and Stroud (2007) and Shen and Huang (2008).

The paper is organized as follows. The functional model and notation are presented in Section 2. The main theoretical results are presented in Section 3, were we first show how to construct the BLUP for the continuation of a curve. Next, we show how the BLUP can be computed efficiently. Confidence bands are discussed in Section 4. In Section 5 we apply the estimator to real-world data, comparing direct and indirect workload forecasting, and comparing our results to other techniques. Concluding remarks appear in Section 6. Technical proofs are provided in the Appendix.

2. The functional model. In this section we present the model and notation that will be used throughout this paper. Let X be a random function defined on the segment $S = [0, T]$, and let the random functions X_1 and X_2 be the restrictions of X to the segments $S_1 = [0, U]$ and $S_2 = [U, T]$, respectively, for some $0 < U < T$. Our goal is to estimate X_2 given information regarding X_1.

We assume that X is of the form

$$X(t) = \mu(t) + \boldsymbol{\phi}(t)'\boldsymbol{h},$$

where $\mu(t)$ is the mean function, $\boldsymbol{h} = (h_1, \ldots, h_p)$ is a random vector with mean zero and covariance matrix L, L is diagonal with $L_{11} \geq \ldots \geq L_{pp} > 0$, and $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_p(t))'$ is a vector of orthonormal functions. We

assume that the functions μ and ϕ_j have a basis expansion with respect to some B-spline basis $\boldsymbol{b} = (b_1, ..., b_N)'$, defined on some fixed knot sequence τ. We denote this B-spline space by $S_{k,\tau}$ where k denotes the the splines' order. Thus, we can write $\mu(t) = \boldsymbol{b}(t)'\boldsymbol{\mu}$ and $\boldsymbol{\phi}(t) = A'\boldsymbol{b}(t)$, for some $p \times 1$ vector $\boldsymbol{\mu}$ and $N \times p$ loading matrix A. Thus, we have

$$(1) \qquad X(t) = \boldsymbol{b}(t)'(\boldsymbol{\mu} + A\boldsymbol{h}) \doteq \boldsymbol{b}(t)'\boldsymbol{x}\,,$$

where $\boldsymbol{x} = \boldsymbol{\mu} + A\boldsymbol{h}$. We think of N, the ambient functional space dimension, as being much larger then p, the dimension of the subspace which spanned by the random function X.

We assume that instead of seeing X, we actually observe some noisy version of X, namely

$$Y(t) = X(t) + \varepsilon(t)\,,$$

where $\varepsilon(t) = \boldsymbol{\psi}(t)'\boldsymbol{\epsilon}$ is some random function independent of $X(t)$, $\boldsymbol{\epsilon}$ is a $q \times 1$ zero-mean random vector with diagonal covariance matrix Σ, and $\boldsymbol{\psi}$ is a vector of functions. Since $X(t)$ is a (random) linear combination of $\phi_1(t), \ldots, \phi_p(t)$, we consider the noise as the part of the observed function $Y(t)$ that cannot be explained using such linear combinations. Hence we assume that $\boldsymbol{\psi}$ is orthogonal to $\boldsymbol{\phi}$. However, note that this orthogonality is not necessarily preserved when $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ are restricted to one of the segments S_1 or S_2. We assume that $\boldsymbol{\psi}$ also has an expansion with respect to the basis $\boldsymbol{b}$ and hence $\boldsymbol{\psi}(t) = B'\boldsymbol{b}(t)$ for some $N \times q$ loading matrix B. Using this notation we may write

$$(2) \qquad Y(t) = \boldsymbol{b}(t)'(\boldsymbol{\mu} + A\boldsymbol{h} + B\boldsymbol{\epsilon})\,.$$

The covariance functions $u(s,t) = \mathrm{Cov}(X(s), X(t))$ and $v(s,t) = \mathrm{Cov}(Y(s), Y(t))$ can be written by $\boldsymbol{b}(s)'(ALA')\boldsymbol{b}(t) \doteq \boldsymbol{b}(s)'g\boldsymbol{b}(t)$ and $\boldsymbol{b}(s)'(ALA'+B\Sigma B')\boldsymbol{b}(t) \doteq \boldsymbol{b}(s)'G\boldsymbol{b}(t)$, respectively. We define the correspondence covariance operators from $S_{k,\tau}$ to itself for functions $f \in S_{k,\tau}$ as

$$
\begin{aligned}
(\gamma f)(t) &= \int_S u(s,t)f(s)\mathrm{d}s = \boldsymbol{b}(t)'gW\boldsymbol{f} \\
(\Gamma f)(t) &= \int_S v(s,t)f(s)\mathrm{d}s = \boldsymbol{b}(t)'GW\boldsymbol{f}
\end{aligned}
$$

where $W = \int_S \boldsymbol{b}(s)\boldsymbol{b}(s)'\mathrm{d}s$, and $\boldsymbol{f}$ is the expansion of the function f in the B-spline basis.

We now introduce the notation for X_1 and X_2 and their respective noisy versions Y_1 and Y_2. Let τ_1 and τ_2 be knot sequences that agree with τ on the segments $[0, U)$ and $(U, T]$, respectively, and have knot multiplicity of k

at U. Let S_{k,τ_i} for $i = 1, 2$ be the k-ordered spline space with knot sequence τ_i and let $\boldsymbol{b_i}(t) = (b_{i1}(t), \ldots, b_{iN_i}(t))$ be its corresponding B-spline basis. We wish to represent X_i and Y_i ($i = 1, 2$) using the representations of X and Y.

Note that when the functions $\mu(t)$, $\phi_j(t)$, $\psi_j(t)$, $v(s,t)$ and $u(s,t)$ are known on $[0, T]$, they are also known on S_1 and S_2. Thus, it is enough to represent these functions using the bases $\boldsymbol{b_i}$ in order to obtain representations for X_i and Y_i. Recall that $\mu(t) = \boldsymbol{b}(t)'\boldsymbol{\mu}$ for some vector of coefficients $\boldsymbol{\mu}$. Using the knot-insertion algorithm (see de Boor, 2001, Chapter 11) we obtain new vectors $\boldsymbol{\mu_i}$ such that (a) $\mu(t) = \boldsymbol{b_i}(t)'\boldsymbol{\mu_i}$ for all t on which $\boldsymbol{b_i}$ is defined and (b) $\boldsymbol{\mu_i}$ is obtained from $\boldsymbol{\mu}$ by truncation and a change of at most k coefficients. Similarly, using the knot-insertion algorithm, we can obtain the loading matrices A_i and B_i such that $\boldsymbol{\phi}(t) = A_i\boldsymbol{b_i}(t)$ and $\boldsymbol{\psi}(t) = B_i\boldsymbol{b_i}(t)$ for all t on which $\boldsymbol{b_i}$ is defined. Summarizing, we have

$$
\begin{aligned}
(3) \qquad X_i(s) &= \boldsymbol{b_i}(s)'(\boldsymbol{\mu_i} + A_i\boldsymbol{h}) \doteq \boldsymbol{b_i}(s)'\boldsymbol{x_i} \\
Y_i(s) &= \boldsymbol{b_i}(s)'(\boldsymbol{\mu_i} + A_i\boldsymbol{h} + B_i\boldsymbol{\epsilon}) \doteq \boldsymbol{b_i}(s)'\boldsymbol{y_i} \\
v(s,t) &= \boldsymbol{b_i}(s)(A_iLA_j' + B_i\Sigma B_j')\boldsymbol{b_j}(t) \doteq \boldsymbol{b_i}(s)'G_{ij}\boldsymbol{b_j}(t) \\
u(s,t) &= \boldsymbol{b_i}(s)(A_iLA_j')\boldsymbol{b_j}(t) \doteq \boldsymbol{b_i}(s)'g_{ij}\boldsymbol{b_j}(t)
\end{aligned}
$$

for $i, j = 1, 2$ and for each $s \in S_i$ and $t \in S_j$.

We define the operators γ_{ij} and Γ_{ij} from S_{k,τ_j} to S_{k,τ_i} for $i, j = 1, 2$ by

$$
\begin{aligned}
(4) \qquad (\gamma_{ij}f)(t) &= \int_{S_j} u(s,t)f(s)\mathrm{d}s = \boldsymbol{b_i}(t)'g_{ij}W_j\boldsymbol{f} \\
(\Gamma_{ij}f)(t) &= \int_{S_j} v(s,t)f(s)\mathrm{d}s = \boldsymbol{b_i}(t)'G_{ij}W_j\boldsymbol{f} \,,
\end{aligned}
$$

where $W_j = \int_{S_j} \boldsymbol{b_j}(s)\boldsymbol{b_j}(s)'\mathrm{d}s$, and $\boldsymbol{f}$ is the expansion of the function f in $\boldsymbol{b_j}$.

The model discussed above will be used for the estimation of X_2 given Y_1. Note that the distributions of X and Y are generally not known. In a realistic situation one needs to estimate the model components. Recall that $Y(t) = \boldsymbol{b}(t)'(\boldsymbol{\mu} + A\boldsymbol{h} + B\boldsymbol{\epsilon})$, where $\boldsymbol{h}$ and $\boldsymbol{\epsilon}$ are random vectors with zero mean and covariance matrices L and Σ, respectively. Before discussing the forecasting procedure, we briefly discuss how estimation of $\boldsymbol{\mu}, L, \Sigma$ and the loading matrices A and B can be performed.

Assume that the functions $Y^{(1)}, \ldots, Y^{(m)}$ were drawn according to the distribution law of Y. We distinguish between two scenarios. In the first scenario we assume that the functions $Y^{(1)}, \ldots, Y^{(m)}$ were observed. In this case one can estimate the various components of Y using functional principal component analysis (functional PCA). This can be done either by using

PCA on the coefficients of the functions or by introducing some smoothness using regularized functional PCA (see, for example, Ramsay and Silverman, 2005, Chapters 8 and 9). The matrices L and Σ are then determined by the eigenvalues of the PCA decomposition while the loading matrices A and B consist of the coefficients of the principal components with respect to the basis $\boldsymbol{b}$. The size of L and Σ can be estimated using the gaps in the eigenvalues of the PCA decomposition.

In the second scenario, we assume that some noisy discrete observations are given; for example in the following form

$$Z^{(i)}(t_{ij}) = Y^{(i)}(t_{ij}) + e_{ij}\,,$$

for $i = 1, \ldots, m$, $j = 1, \ldots, n_j$, and $0 \le t_{i1} < \ldots < t_{in_j} \le T$, and where $e_{ij} \sim N(0, \sigma^2)$ are independent. In this case, one can first estimate the functions and then use functional PCA as described above. The simplest way to estimate the functions is to estimate each function separately, using, for example, regression splines (de Boor, 2001, Chapter 14). This method is used in the numerical examples in Section 5. Others, such as Kneip (1994) and Besse, Cardot and Ferraty (1997), suggest to estimate all the functions simultaneously. Both methods use some sort of functional PCA. These methods suggest ways to estimate the length of $\boldsymbol{h}$. The method by Besse, Cardot and Ferraty (1997) also assumes a splines environment, as in our case.

3. The construction of the BLUP. Given Y_1, the noisy version of the beginning part of the random function X, our goal is to find a *good* estimator for X_2, the continuation of X_1.

Following Robinson (1991), we say that $\hat{X}_2$ is a *good* estimator of X_2 given Y_1 if the following criteria hold:

(C1) $\hat{X}_2$ is a linear function of Y_1.
(C2) $\hat{X}_2$ is unbiased, i.e., $E[\hat{X}_2(t)] = \mu(t)$.
(C3) $\hat{X}_2$ has minimum mean square error among the class of linear unbiased estimators.

Two more demands regarding the estimator that seems desirable in our context are

(C4) The random function $\hat{X}_2$ lies in the space S_{k,τ_2}.
(C5) When no noise is introduced, i.e., when $Y_1 = X_1$, the concatenation of $\hat{X}_2$ to X_1 lies in $S_{k,\tau}$; in other words, the combined function

$$\hat{X} = \left\{ \begin{array}{ll} X_1(t) & 0 \le t \le U \\ \hat{X}_2(t) & U < t \le T \end{array} \right.$$

is smooth enough.

An estimator that fulfills (C1)-(C5) will be referred to as a best linear unbiased predictor (BLUP). In this section we will show how to construct such a BLUP and prove that is is defined uniquely.

REMARK 3.1.   *Note that the definition of unbiased estimator in (C2) is not the usual definition. A more restrictive criterion is*

(C2*) $\hat{X}_2$ *is unbiased in the the following sense* $E[\hat{X}_2(t)|Y_1] = E[X_2(t)|Y_1]$.

We will show that when Y *is a Gaussian process, this criterion is fulfilled by the proposed BLUP as well.*

REMARK 3.2.   *The analogous results in the multivariate case are well known. Here* best *estimator means estimator that meets criteria (C1)-(C3). Let* $Z = (Z_1, Z_2)'$ *be a random vector such that*

$$E \left[\begin{array}{c} Z_1 \\ Z_2 \end{array} \right] = \left(\begin{array}{c} m_1 \\ m_2 \end{array} \right), \quad \mathrm{Var} \left[\begin{array}{c} Z_1 \\ Z_2 \end{array} \right] = R = \left(\begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array} \right).$$

Then the BLUP of $Z_2|Z_1$ *is given by*

(5) $$\hat{Z}_2 = m_2 + R_{21} R_{11}^+ (Z_1 - m_1)$$

where R_{11}^+ *is the Moore-Penrose pseudoinverse of* R_{11} *(see, for example, Marsaglia, 1964).*

In the following, we define the linear operators that are the analogs of the matrices R_{11}^+ and R_{21} from the multivariate case. This enables the construction of a uniquely-defined BLUP for X_2.

We begin with defining the operator $\Gamma_{11}^+ : S_{k,\tau_1} \rightarrow S_{k,\tau_1}$, which is the functional equivalent of R_{11}^+. Define the function

$$v_{11}^+(s,t) = \boldsymbol{b_1}(s)' W_1^{-1} G_{11}^+ W_1^{-1} \boldsymbol{b_1}(t),$$

for every $s, t \in S_1$. Note that W_1 is invertible since it is a Gram matrix of basis functions (see Sansone, 1991, Theorem 1.5). Define

$$(\Gamma_{11}^+ f)(t) = \int_{S_1} v_{11}^+(s,t) f(s) \mathrm{d}s = \boldsymbol{b_1}(t)' W_1^{-1} G_{11}^+ \boldsymbol{f},$$

where $\boldsymbol{f}$ is the expansion of the function f in the B-spline basis $\boldsymbol{b_1}$. The following lemma justifies the notation of Γ_{11}^+ as a pseudoinverse operator.

LEMMA 3.3. *With probability one,*

$$\Gamma_{11}\Gamma_{11}^{+}(Y_1 - \mu) = \Gamma_{11}^{+}\Gamma_{11}(Y_1 - \mu) = Y_1 - \mu\,.$$

See proof in the Appendix.

We are now ready to define the estimator for X_2 given Y_1, similarly to estimator (5) in the multivariate case, by

$$(6) \quad \hat{X}_2(t) = \mu(t) + \gamma_{21}\Gamma_{11}^{+}(Y_1 - \mu)(t) = \boldsymbol{b_2}(t)'(\boldsymbol{\mu_2} + g_{21}G_{11}^{+}(\boldsymbol{y_1} - \boldsymbol{\mu_1}))\,,$$

for every $t \in S_2$. Then we have

THEOREM 3.4. *The estimator $\hat{X}_2$ meets criteria (C1)-(C5) and is unique up to equivalence. Moreover, if Y is a Gaussian process, then $\hat{X}_2$ meets criterion (C2*) as well.*

PROOF. We show that (C1)-(C5) hold, one by one.

(C1) holds because $\hat{X}_2$ is indeed a linear transformation of Y_1 as can be seen from (6).

(C2) holds since

$$E[\hat{X}_2(t)] = \boldsymbol{b_2}(t)'(\boldsymbol{\mu_2} + g_{21}G_{11}^{+}(E[\boldsymbol{y_1} - \boldsymbol{\mu_1}])) = \boldsymbol{b_2}(t)'\boldsymbol{\mu_2} = \mu(t)\,.$$

(C3) states that $\hat{X}_2$ should minimize the mean square error among all the unbiased linear estimators . Let $\tilde{X}_2$ be another linear unbiased estimator. Then we can write $\tilde{X}_2 = (\tilde{X}_2 - \hat{X}_2) + \hat{X}_2$. Since both $\tilde{X}_2$ and $\hat{X}_2$ are unbiased, $\tilde{X}_2 - \hat{X}_2$ is an unbiased linear estimator of zero, hence it is of the form $\boldsymbol{b_2}(t)'M(\boldsymbol{y_1} - \boldsymbol{\mu_1})$ for some $N_2 \times N_1$ matrix M. Moreover, it can be shown that $\text{Cov}(X_2 - \hat{X}_2, \tilde{X}_2 - \hat{X}_2) = 0$. Indeed,

$$
\begin{aligned}
\text{Cov}\big((X_2 - \hat{X}_2)(s), (\tilde{X}_2 - \hat{X}_2)(t)\big) &= E[(X_2 - \hat{X}_2)(\tilde{X}_2 - \hat{X}_2)(t)] \\
&= \boldsymbol{b_2}(s)'E[(\boldsymbol{x_2} - \boldsymbol{\mu_2})(\boldsymbol{y_1} - \boldsymbol{\mu_1})']M'\boldsymbol{b_2}(t) \\
&\quad -\boldsymbol{b_2}(s)'E[\boldsymbol{\mu_2} + g_{21}G_{11}^{+}(\boldsymbol{y_1} - \boldsymbol{\mu_1})(\boldsymbol{y_1} - \boldsymbol{\mu_1})']M'\boldsymbol{b_2}(t) \\
&= \boldsymbol{b_2}(s)'(g_{21}M' + g_{21}G_{11}^{+}G_{11}M'))\boldsymbol{b_2}(t) = 0\,.
\end{aligned}
$$

where the last equality follows from Lemma 3.3.

To see that $\hat{X}_2$ minimizes the mean square error, note that

$$
\begin{aligned}
E[(X_2 - \tilde{X}_2)^2(t)] &= E[(X_2 - \hat{X}_2)^2(t)] + E[(\tilde{X}_2 - \hat{X}_2)^2(t)] + 2E[(X_2 - \hat{X}_2)(\hat{X}_2 - \tilde{X}_2)(t)] \\
(7) &= E[(X_2 - \hat{X}_2)^2(t)] + E[(\tilde{X}_2 - \hat{X}_2)^2(t)] \geq E[(X_2 - \hat{X}_2)^2(t)]\,,
\end{aligned}
$$

which proves that $\hat{X}_2$ minimizes the mean square error and is unique up to equivalence.

(C4) holds by construction.

(C5) states that when no noise is introduced, $\hat{X}_2$ a smooth continuation of X_1. First, note that by Lemma 3.3

$$X_1(t) = \boldsymbol{b_1}(t)'(\boldsymbol{\mu_1} + G_{11}G_{11}^+(\boldsymbol{x_1} - \boldsymbol{\mu_1})) = \boldsymbol{b_1}(t)'(\boldsymbol{\mu_1} + A_1(LA_1'G_{11}^+)(\boldsymbol{x_1} - \boldsymbol{\mu_1})).$$

By definition we also have

$$\hat{X}_2(t) = \boldsymbol{b_2}(t)'(\boldsymbol{\mu_2} + g_{21}G_{11}^+(\boldsymbol{x_1} - \boldsymbol{\mu_1})) = \boldsymbol{b_2}(t)'(\boldsymbol{\mu_2} + A_2(LA_1'G_{11}^+)(\boldsymbol{x_1} - \boldsymbol{\mu_1})).$$

Define $\hat{X}(t) = \boldsymbol{b}(t)'(\mu(t) + A(LA_1'G_{11}^+)(\boldsymbol{x_1} - \boldsymbol{\mu_1}))$. It follows from the definitions of $\boldsymbol{\mu_i}$, A_i and $\boldsymbol{b_i}$ that $\hat{X}(t)$ agrees with X_1 on S_1 and with $\hat{X}_2$ on S_2. Since $\hat{X} \in S_{k,\tau}$, the result follows.

Finally, if Y is a Gaussian process, then $\boldsymbol{y_1}$ and $\boldsymbol{x_2}$ are normally distributed such that $\mathrm{Var}(\boldsymbol{y_1}) = G_{11}$ and $\mathrm{Cov}(\boldsymbol{x_2}, \boldsymbol{y_1}) = g_{21}$. Following Marsaglia (1964) we obtain

$$
\begin{aligned}
(8) \quad E[X_2(t)|Y_1] &= \boldsymbol{b}(t)'E[\boldsymbol{x_2}|\boldsymbol{y_1}] = \boldsymbol{b}(t)'(\boldsymbol{\mu_2} + g_{21}G_{11}^+(\boldsymbol{y_1} - \boldsymbol{\mu_1})) \\
&= \hat{X}_2(t) = E[\hat{X}_2(t)|Y_1]
\end{aligned}
$$

and criterion (C2*) is met. $\square$

It should be noted that when the parameters of the model are estimated (see end of Section 2) and a Gaussian model is assumed, the estimator $\hat{X}_2$ can be considered as an empirical Bayes estimator. Indeed, the estimation of the distribution of $\boldsymbol{h}$ and $\boldsymbol{\epsilon}$ can be considered as estimating the prior distribution, while the the computation of $\hat{X}_2$ as in (8) is in fact finding the posterior mean given the data Y_1.

From a computational point of view, the computation of $\hat{X}_2$ may seem heavy. Indeed by (6) it involves finding the pseudoinverse of G_{11}^+ which is an $N_1 \times N_1$ matrix. However, a simpler expression can be found. Recall that

$$G_{11} = [A_1, B_1]\begin{bmatrix} L & 0 \\ 0 & \Sigma \end{bmatrix}\begin{bmatrix} A_1' \\ B_1' \end{bmatrix} \doteq CSC'.$$

where $C = [A_1, B_1]$ and $S = \begin{bmatrix} L & 0 \\ 0 & \Sigma \end{bmatrix}$. Using Lemma A.1.3 with $T = S^{1/2}C'$ we have

$$
\begin{aligned}
G_{11}^+ &= CS^{1/2}\left(\left(S^{1/2}C'CS^{1/2}\right)^+\right)^2 S^{1/2}C' \\
&= CS^{1/2}\left(S^{-1/2}(C'C)^+S^{-1}(C'C)^+S^{-1/2}\right)S^{1/2}C' \\
&= C(C'C)^+S^{-1}(C'C)^+C',
\end{aligned}
$$

which involves the pseudoinverse computation of a $(p + q) \times (p + q)$ matrix.

Finally, instead of assuming that $Y_1(t) = \boldsymbol{b_1}(t)'(\boldsymbol{\mu_1} + A_1\boldsymbol{h} + B_1\boldsymbol{\epsilon})$, one may assume that

$$Y_1(t) = \boldsymbol{b_1}(t)'(\boldsymbol{\mu_1} + A_1\boldsymbol{h} + \tilde{\boldsymbol{\varepsilon}}_1)$$

where $\tilde{\boldsymbol{\varepsilon}}_1$ is a $N_1 \times 1$ mean zero random vector with $\sigma^2 I$ covariance matrix and I is the identity matrix. In this case,

$$(9) \qquad \hat{X}_2(t) = \boldsymbol{b_2}(t)'(\boldsymbol{\mu_2} + g_{21}(A_1 L A_1' + \sigma^2 I)^{-1}(\boldsymbol{x_1} - \boldsymbol{\mu_1}))$$

which is the ridge regression estimator (Hoerl and Kennard, 1970). Once again, a simpler expression can be obtained using some matrix algebra (see Robinson, 1991, Eq. 5.2). We have

$$g_{21}(A_1 L A_1' + \sigma^2 I)^{-1} = A_2 L A_1'(A_1 L A_1' + \sigma^2 I)^{-1} = A_2 \left(A_1' A_1 + \sigma^2 L^{-1}\right)^{-1} A_1',$$

and hence $\hat{X}_2(t) = \boldsymbol{b_2}(t)'(\boldsymbol{\mu_2} + A_2 \left(A_1' A_1 + \sigma^2 L^{-1}\right)^{-1} A_1'(\boldsymbol{x_1} - \boldsymbol{\mu_1}))$, which involves only the inverse of a $p \times p$ matrix.

4. Confidence Bands. In Section 3 we suggested the estimator $\hat{X}_2$ for the continuation of the function X_1. In this section we would like to construct confidence bands for this estimator. We consider two kinds of confidence bands. The first is a global confidence band. A global confidence band with confidence level $(1 - \delta)100\%$ is defined as a pair of functions, the upper band f_U and the lower band f_L, such that $P(f_L(t) < X_2(t) < f_U(t)$ for all $t \in S_2) \geq 1 - \delta$. We also consider local confidence bands. Local confidence bands do not require that the last condition holds simultaneously for all t; rather we are looking for a pair of functions g_U and g_L such that for all $t \in S_2$, $P(g_L(t) < X_2(t) < g_U(t)) \geq 1 - \delta$.

Our construction of both global and local confidence bands is based on the technique introduced by Knafl, Sacks and Ylvisaker (1985). The idea is the following. We first create simultaneous confidence intervals for some finite set of point. Then, using the attributes of spline functions, we complete this band for all points of S_2. The computation of these bands can be computationally demanding. Hence, we suggest also confidence bands that are based on cross-validation. While these confidence bands do not have the theoretical guarantee of the former, they are simple to compute and seem to work reasonably well (see Section 5, Table 4).

In the following, we assume that X and Y are Gaussian processes. Therefore X_2 is also a Gaussian process and, by (8), $E[X_2|Y_1] = \hat{X}_2$. Similarly, we have

$$\text{Cov}(X_2(s), X_2(t)|Y_1) = \boldsymbol{b_2}(s)' \left(g_{22} - G_{21} G_{11}^+ G_{12}\right) \boldsymbol{b_2}(t).$$

Define

$$Z(t) = \frac{X_2(t) - \hat{X}_2(t)}{\text{Var}(X_2(t)|Y_1)^{1/2}} \, ,$$

then $Z(t)$ is a zero-mean Gaussian process with variance 1 for each t.

Let $t_1, \ldots, t_m$ be the breaks in τ_2, i.e., the knots of τ_2, ignoring knot multiplicity. Let $t_{i,j} = t_i + \frac{j-1}{k-1}(t_{i+1} - t_i)$, $j = 1, \ldots, k-1$. Define the following grid

$$\mathcal{G} = \{t_{1,1}, t_{1,2}, \ldots, t_{m-1,k-1}, t_m\} \, ,$$

i.e., $\mathcal{G}$ is a grid that includes all the breaks in τ_2 and there are $k-2$ equally spaced grid points between each two successive breaks of τ_2. We are interested in computing simultaneous confidence intervals for the points in $\mathcal{G}$. In other words, for a given δ, we would like to find z_δ such that
(10)
$$P(\max_{t \in \mathcal{G}} |X_2(t) - \hat{X}_2(t)| > z_\delta \text{Var}(X_2(t)|Y_1)^{1/2}) = P(\max_{t \in \mathcal{G}} |Z(t)| > z_\delta) \le \delta \, .$$

z_δ can be found using simulations or by utilizing the inequality (Knafl, Sacks and Ylvisaker, 1985, Eq. (1.8))
(11)
$$P(\max_{t \in \mathcal{G}} |Z(t)| > a) \le P(|Z(t_{1,1})| > a) + \sum_{i=1}^{m-1} \sum_{j=1}^{k-1} P(|Z(t_{i,j})| \le a \, , |Z(t_{i,j+1})| > a) \, .$$

Recall that the trajectories of $(X_2(t) - \hat{X}_2)|Y_1$ are in S_{k,τ_2}. Hence for each segment between two successive breaks of τ_2, say $[t_i, t_{i+1}]$, the trajectories are k-ordered polynomials. Let $p(t)$ be a restriction of such a trajectory to $[t_i, t_{i+1}]$. $p(t)$ can be written, using Lagrange polynomials, as

$$p(t) = \sum_{j=1}^k \ell(t) p(t_{i,j}) \quad ; \quad \ell_j(t) = \prod_{r=1, \, r \neq j}^k \frac{t - t_{i,r}}{t_{i,j} - t_{i,r}} \, .$$

Note that for all $t \in [t_i, t_{i+1}]$, $|p(t)| \le \sum_{r=1}^k |\ell(t)| p(t_{i,j})$. Hence, if

(12) $$|p(t_{i,j})| < z_\delta \text{Var}(X_2(t_{i,j})|Y_1)^{1/2}) \qquad \text{for } j = 1, \ldots, k \, ,$$

then for all $t \in [t_i, t_{i+1}]$

(13) $$|p(t)| < z_\delta \sum_{j=1}^k |\ell_j(t)| \text{Var}(X_2(t_{i,j})|Y_1)^{1/2}) \doteq z_\delta D_{t_i}(t) \, .$$

By (10) we have that with probability greater than or equal to $1 - \delta$, the inequality in (12) holds simultaneously for all i. Thus, with probability greater

than or equal to $1 - \delta$, the inequality in (13) also holds. Define the pair of functions (f_U, f_L) on S_2 such that for all $t \in [t_i, t_{i+1}]$

$$(14) \qquad f_U(t) = \hat{X}_2(t) + z_\delta D_{t_i}(t) \quad ; \quad f_L(t) = \hat{X}_2(t) - z_\delta D_{t_i}(t) \,.$$

Then (f_U, f_L) are global confidence band for $X_2|Y_1$ with a confidence level greater than or equal to $100(1 - \delta)\%$. Note that f_U and f_L are continuous.

For local confidence bands, we can define the pair of functions (g_U, g_L) on S_2 such that for all $t \in [t_i, t_{i+1}]$

$$(15) \qquad g_U(t) = \hat{X}_2(t) + \hat{z}_\delta D_{t_i}(t) \quad ; \quad g_L(t) = \hat{X}_2(t) - \hat{z}_\delta D_{t_i}(t) \,,$$

where

$$\hat{z}_\delta = \max_i \min \left\{ z_\delta : P\left(\max_{j=1,\ldots,k} |Z(t_{i,j})| > z_\delta \right) \leq \delta \right\} \,.$$

Using $\hat{z}_\delta$ ensures that g_U and g_L are continuous. The estimation of $\hat{z}_\delta$ can be done using the relation in (11). We note that in the computation of $\hat{z}_\delta$ we demanded that between each two successive breaks in τ_2, with probability greater than $1 - \delta$ the trajectories of X_2 will stay within the band. While this can be restrictive if the distance between successive points in τ_2 is large, a simple solution is to take the set $\mathcal{G}$ to be more dense.

We remark here on some issues related to the confidence bands defined in (14-15). First, note that the bands are conservative, meaning that the confidence level is greater than $100(1-\delta)\%$. Second, we have assumed that $X_2|Y_1$ is a Gaussian process with known distribution. Third, the computation of z_δ (or $\hat{z}_\delta$) can be demanding. Hence, we suggest to estimate confidence bands from the data using some sort of cross-validation. Compute $\mathrm{Var}(X_2(t)|Y_1)^{1/2}$ for all $t \in \mathcal{G}$, and let $\hat{D}(t)$ be the k-ordered regression spline function with knot sequence τ_2 of the points $\{(t, \mathrm{Var}(X_2(t)|Y_1)^{1/2}) : t \in \mathcal{G}\}$. We suggest the following confidence bands

$$(16) \qquad \hat{f}_U(t) = \hat{X}_2(t) + C_{\mathrm{Global}}\hat{D}(t) \quad ; \quad \hat{g}_U(t) = \hat{X}_2(t) + C_{\mathrm{Local}}\hat{D}(t) \,,$$

and similarly for $\hat{f}_L$ and $\hat{g}_L$ where C_{Global} and C_{Local} are computed using cross-validation as described below. Assume that the functions $Y^{(1)}, \ldots, Y^{(m)}$ were observed. Partition the functions to K folds $F_j : j = 1, \ldots, K$. Compute $\hat{X}_2(t)$ and $\hat{D}(t)$ for each subset of $K - 1$ folds. Define

$$C_{\mathrm{Global},j} = \min \left\{ c > 0 : \frac{1}{|F_j|} \sum_{Y_i \in F_j} I\{|Y_i(t) - \hat{X}_2(t)| < c\hat{D}(t) \text{ for all } t \in \mathcal{G}\} > 1 - \delta \right\}$$

$$C_{\mathrm{Local},j} = \min \left\{ c > 0 : \min_{t \in \mathcal{G}} \left(\frac{1}{|F_j|} \sum_{Y_i \in F_j} I\{|Y_i(t) - \hat{X}_2(t)| < c\hat{D}(t)\} \right) > 1 - \delta \right\}$$

where $I\{B\}$ is the indicator function of the set B. Then we suggest to choose C_{Global} and C_{Local} to be the median of $C_{\text{Global},j}$ and $C_{\text{Local},j}$ respectively. We note that the suggestion to extend the confidence bands from points in the grid to the whole segment using regression splines seems reasonable when the grid is fine enough. In the numerical examples of Section 5 we compute the confidence bands using the cross-validation technique.

5. Numerical Examples. In this section we apply the estimator $\hat{X}_2$ to call center data. We are interested in forecasting the continuation of two processes: the arrival process and the workload process. The estimators of these two processes play an important roll in determining staffing level at call centers (see, for example, Aldor-Noiman, Feigin and Mandelbaum, 2009; Shen and Huang, 2008; Reich, 2010). Usually, staffing levels are determined in advance, at least one day ahead. Here we propose a method for updating the staffing level, given information obtained throughout the beginning of the day. As noted by Gans, Koole and Mandelbaum (2003) and by Shen and Huang (2008), such updating is operationally beneficial and feasible. If performed appropriately, it could result in higher efficiency and service quality: based on the revised forecasts, a manager can adjust staffing levels correspondingly, by offering overtime to agents on duty or dismissing agents early, calling in additional agents if needed, increasing or reducing cross-selling, and transferring agents to other activities such as email inquiries and faxes.

This section is organized as follows. We first describe the arrival and workload processes (Section 5.1). We then describe the data (Section 5.2) and the forecast implementation (Section 5.3). The analysis appears in Sections 5.4-5.6. Finally, confidence bands are discussed in Section 5.7.

5.1. *The arrival and workload processes.* We define the arrival process of day j, $a_j(t)$, as the number of calls that arrive on day j during the time interval $[t-c, t]$, where t varies continuously over time and c is some fixed constant. Note that $a_j(t)$ itself is not a continuous function, but when the call volume is large and this function does not change drastically over short time intervals, it can be assumed that the function $a_j(t)$, for each day j, arises from some underlying deterministic smooth arrival rate function $\lambda(t)$ plus some noise (Weinberg, Brown and Stroud, 2007). In this case $a_j(t)/c$ can be considered as an approximation of the smooth function $\lambda(t)$. We now describe the workload process $w_j(t)$ for each day j. The function $w_j(t)$ counts the number of calls that would have been handled by the call center on day j at time t, *assuming an unlimited number of agents and hence no abandonments.* From a management point of view, the advantage of looking

at $w_j(t)$ over looking at $a_j(t)$ is that $w_j(t)$ reflects the number of agents actually needed at each point in time. However, as opposed to the process $a_j(t)$, which is observable in real time, the computation of $w_j(t)$, for a specific time t, involves estimation of call durations for abandoned calls and can be performed only after all calls entered up to time t are actuality served (see the discussion at Aldor-Noiman, Feigin and Mandelbaum, 2009; Reich, 2010).

5.2. *The data.* The data used for the forecasting examples were gathered at a call center of a large U.S. commercial bank. The bank has various types of operations such as retail banking, consumer lending, and private banking. Since the call arrival pattern varies over different types of services, we restrict attention to retail services, which account for approximately 70% of the calls (see Weinberg, Brown and Stroud, 2007). The first two examples are of the arrival process and the workload process, for weekdays between March and October 2003. The data for the first example consists of the arrival counts at five-minutes resolution between 7:00 AM and 9:05 PM (i.e., $c = 5$ in the definition of $a_j(t)$). The data for the second example consists of average workload, also in five-minutes resolution, between 7:00 AM and 9:05 PM. There are 164 days in the data set after excluding some abnormal days such as holidays. Figure 1 shows arrival count profiles for different days of the week.

The third example explores the arrival process during weekends between March and October 2003. There are 67 days in the data set (excluding one day with incomplete data). As can be seen from Figure 1, the weekend behavior is different from that of the working days, and there is a Saturday pattern and a Sunday pattern. The data for this example consists of the arrival counts at fifteen-minutes resolution between 8 AM and 5 PM. The change in interval length from the previous two examples is due to the decreased call-counts. The change in day length is due to the low activity in early morning and late afternoon hours on weekends (see Figure 1).

In the first and second examples, we used the first 100 weekdays as the training set and the last 64 weekdays as the test set. For each day from day 101 to day 164, we extracted the *same-weekday* information from the preceding 100 days. Thus, for each day of the week we have about 20 training days. For the third example, the test set consists of weekend days 41 to 67 while the training set for each day consist of its previous 40 weekend days. Thus, similarly, for each day we have about 20 training days. Additionally, we used the data from day start, up to 10 AM and up to 12 PM. All forecasts were evaluated using the data after 12 PM, which enabled fair comparison between the results of the different cut points (10 AM and 12 PM). We also
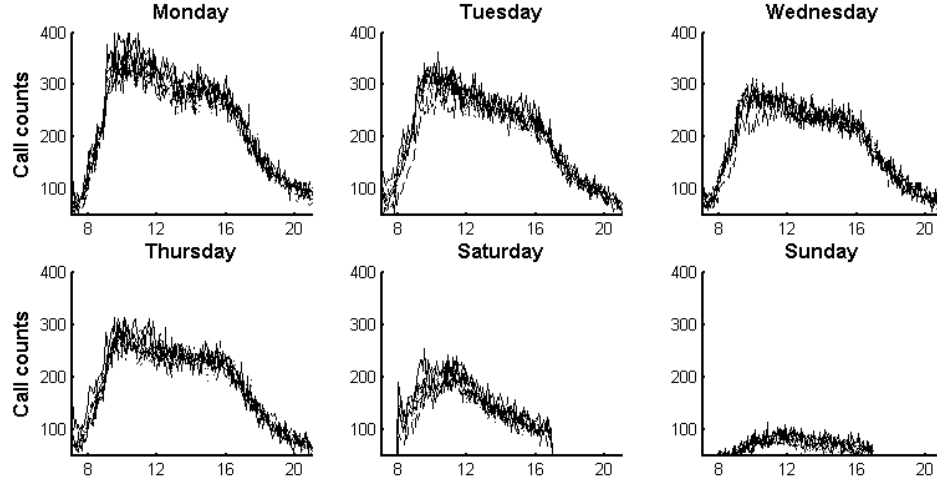
FIG 1. *Arrival count in five-minutes resolution for six successive weeks, grouped according to weekday (Friday was omitted due to space constraints). There is a clear difference between workdays, Saturdays, and Sundays. For the working days, it seems that there is some common pattern. Between 7 AM and 10 AM the call count rises sharply to its peak. Then it decreases gradually until 4 PM. From 4 PM to 5 PM there is a rapid decrease followed by a more gradual decrease from 5 PM until 12 AM. The call counts are smaller for Saturday and much smaller for Sunday. Note also that the main activity hours for weekends are 8 AM to 5 PM, as expected.*

compare our results to the mean of the preceding days, from 12 PM on.

For a detailed description of the first example's data, the reader is referred to Weinberg, Brown and Stroud (2007), Section 2. For an explanation of how the second example's workload process was computed, the reader is referred to Reich (2010). The data for the third example was extracted using SEEStat, which is a software written at the Technion SEELab[1]. We refer the reader to Donin *et al.* (2006) for a detailed description of the U.S. commercial bank call-center data from which the data for all three examples was extracted. The U.S. bank call-center data is publicly downloaded from SEESLab server[1].

5.3. *Forecast implementation.* The forecast was performed by Matlab implementation of the BLUP algorithm from Section 3, where we enable regularization as in (9). For the implementation we used the functional data

---

[1]SEELab: The Technion Laboratory for Service Enterprise Engineering. Webpage: http://ie.technion.ac.il/Labs/Serveng

analysis Matlab library, written by Ramsay and Silverman[2]. The Matlab code, as well as the data sets, are downloadable (see Supplement A). The parameters for the forecast were chosen using 10-fold cross-validation (see end of Section 2). We computed local confidence bands with 95% confidence level using cross-validation, as described in (16). We quantified the results using both Root Mean Squared Error (RMSE) and Average Percent Error (APE), which are defined as follows. For each day j, let

$$ RMSE_j = \left(\frac{1}{K} \sum_{k=1}^{K} (N_{jk} - \hat{N}_{jk})^2 \right)^{1/2} \quad ; \quad APE_j = \frac{100}{K} \sum_{k=1}^{K} \frac{|N_{jk} - \hat{N}_{jk}|}{N_{jk}} , $$

where N_{jk} is the actual number of calls (mean workload) at the k-th time interval of day j in the arrival (workload) process application, $\hat{N}_{jk}$ is the forecast of N_{jk}, and K is the number of intervals.

5.4. *First example: Arrival process for weekdays data.* Forecasting the arrival process for the first example data was studied by both Weinberg, Brown and Stroud (2007) and Shen and Huang (2008). Weinberg, Brown and Stroud assumed that the day patterns behave according to an autoregressive model. The algorithm they suggest first gives a forecast for the current day based on previous days' data. The algorithm estimates the parameters in the autoregressive model using Bayesian techniques. An update for the continuation of the current day forecast is obtained by conditioning on the data of the current day up to the cut point. We refer to this algorithm as Bayesian update (BU) for short. Similarly, the algorithm by Shen and Huang assumes an autoregressive model and gives a forecast for the current day. They then update this forecast using least-square penalization, assuming an underlying discrete process. We will refer to this algorithm as penalized least square (PLS).

Comparison between the results of all three algorithms for the first data set appears in Table 1. Note that for all of the algorithms and all of the categories there is improvement in the 10 AM and 12 PM forecasts over the forecast based solely on past days. The RMSE mean decreases by about 5-13% for the 10 AM forecast, and by 12-15% for the 12 PM forecast, depending on the algorithm. It should be noted that the algorithms by Weinberg, Brown and Stroud and by Shen and Huang use information from all 100 previous days and the knowledge of the previous day call counts. In comparison, the BLUP algorithm uses only the same weekday information ($\sim$20 days) and

---

[2]The functional data analysis Matlab library can be download form ftp://ego.psych. mcgill.ca/pub/ramsay/FDAfuns/Matlab/

the previous day information is not part of its training set. Nevertheless, the results are similar.

The forecasting results for the week that follows Labor Day appear in Figure 2. It can be seen that for the Tuesday that follows Labor Day (Monday) the call counts are much higher than usual. This is captured, to some degree, by the 10 AM forecast and much better by the 12 PM forecast. The same phenomenon occurs, with less strength, during the Wednesday and Thursday following Labor Day, until on Friday all the forecasts become roughly the same. It seems that the power of the continuation-of-curve forecasting is exactly in such situations, in which the call counts are substantially different than usual throughout the day, due to either predictable events, such as holidays, or unpredictable events.

5.5. *Second example: Workload process for weekdays data.* The second example consists of the workload process for weekdays data for the same period as the first example. We forecast the workload process based on these sets of data: previous days' data, up to 10 AM data, and up to 12 PM data. We refer to this forecast as *direct workload forecast since we use past workload estimation as the basis for the forecast. An alternative (and simpler) workload forecasting method was proposed by Aldor-Noiman, Feigin and Mandelbaum (2009). Aldor-Noiman, Feigin and Mandelbaum suggest to forecast the workload by multiplying the forecasted arrival rate by the estimated average service time (see Aldor-Noiman, Feigin and Mandelbaum, 2009, Eq. 21). We refer to this method as* indirect workload forecasting.

Comparison between the two methods appears in Table 2. Following Aldor-Noiman, Feigin and Mandelbaum (2009), we estimated the average service time over a 30 minute period for indirect workload computations.

Example 1	Previous day	10 AM			12 PM		
RMSE	mean	BU	PLS	BLUP	BU	PLS	BLUP
Minimum	12.46	**11.08**		11.51	**11.07**		11.51
Q1	14.11	14.00	**13.31**	13.51	13.56	13.33	**13.27**
Median	16.40	15.50	14.87	**14.69**	14.80	14.60	**14.17**
Mean	19.11	17.86	**16.48**	16.83	16.59	**16.13**	16.15
Q3	21.27	19.87	17.26	**17.04**	16.58	16.39	**15.92**
Maximum	68.93	57.72		**52.09**	53.66		**51.03**

TABLE 1. *Summary of statistics (minimum, lower quartile (Q1), median, mean, upper quartile (Q3), maximum) of RMSE for the forecast based on the mean of the previous days, and BU, PLS, and BLUP using data up to 10 AM and up to 12 PM for the call arrival data set. The results for BU and PLS were taken from the original papers. No maximum and minimum results were given for PLS.*

Note that the direct workload forecast results are slightly better than the indirect workload forecast in most of the categories. Also note that in almost all categories, there is an improvement in the 10 AM and 12 PM forecasts over the forecast based solely on past days. The RMSE mean decreases by about 11% (9%) for the 10 AM forecast, and by 15% (12%) for the 12 PM forecast for the direct (indirect) forecast. Figure 3 presents a visual comparison between the direct and the indirect forecast methods on a specific day. The two forecasts look roughly the same, which is also true for all other days in this data set.

While in this example there is no significant difference between the direct and indirect workload forecasts, we expect these methods to obtain different forecasts when the arrival rate changes during an average service time. This
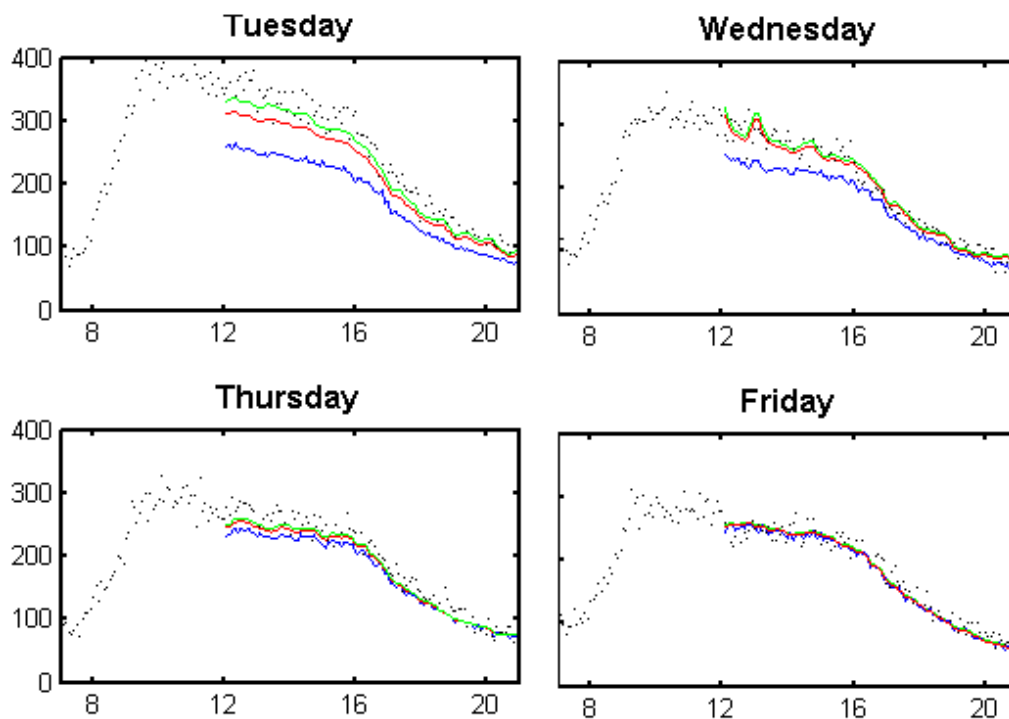

Fig 2. *Forecasting results for the week following Labor Day (Sept. 2-5, 2003) for the call arrival process of the first example. Labor Day itself (Monday) does not appear since holiday data is not included in the data set. The black dots represent the true call counts in five-minutes resolution. The forecasts based on previous days, 10 AM data, and 12 PM data are represented by the blue, red, and green lines, respectively.*

is true, for example, for arrival and service of patients in emergency rooms. The arrival rates of patients to emergency rooms can change within an hour while the time that a patient spends in emergency room (the "service time") is typically on the order of hours. As pointed out by Rozenshmidt (2008, Section 6), in such cases, forecasting the workload by the arrival count multiplied by the average service time may not be accurate. This is because the number of customers in the system is cumulative, while the arrival count counts only those who arrive in the current time interval. Thus, if the arrival count is lower than it was in the previous time interval and the average service time is long, the workload is underestimated. Similarly, if the arrival count is larger than previously, the workload is overestimated.

5.6. *Third example: Arrival process for weekends data.* The third example it that of the weekends' arrivals. The main difference between the first two examples and this one is that the data in this example cannot be considered as data from successive days, due the six day difference between any Sunday and it successive Saturday. Recall that the models considered by Weinberg, Brown and Stroud (2007) and Shen and Huang (2008) have an autoregressive structure. Since this autoregressive structure seems not to hold for this data, the techniques by Weinberg, Brown and Stroud and Shen and Huang are not directly applicable. But even when the autoregressive structure does not hold, the results appearing in Table 3 reveal that forecasting for this data set is still beneficial. Indeed, the RMSE (APE) mean decreases by about 5% (2%) for the 10 AM forecast, and by 10% (4%) for the 12 PM forecast. While these results are not as good as the results in the previous examples, note that the curves in this example begin an hour later and while the call counts are lower during weekends, the arrival rate

Example 2	Day ahead		10 AM		12 PM	
RMSE	Workload (indirect)	Workload (direct)	Workload (indirect)	Workload (direct)	Workload (indirect)	Workload (direct)
Minimum	8.72	8.41	7.98	**7.71**	7.96	8.50
Q1	10.76	10.58	10.21	10.27	10.21	**10.11**
Median	12.10	12.26	11.63	11.21	11.66	**11.05**
Mean	15.97	15.95	14.59	14.26	14.13	**13.48**
Q3	15.08	15.21	14.53	14.20	13.89	**12.76**
Maximum	96.09	94.79	95.74	85.11	93.39	**71.20**

TABLE 2. *Summary of statistics (minimum, lower quartile (Q1), median, mean, upper quartile (Q3), maximum) of RMSE for the forecast based on the mean of the previous days' data, up to 10 AM data and up to 12 PM data, for the workload data set, for both the indirect and the direct forecast methods using the BLUP.*
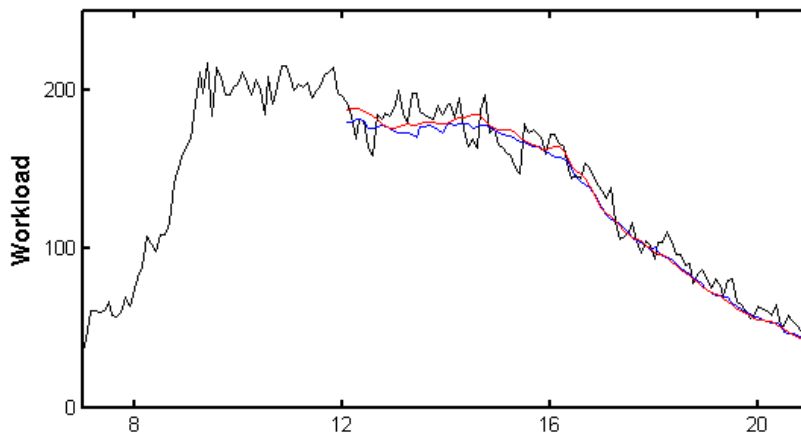
FIG 3. *Workload forecasting for Friday, September 5, 2003, using both the direct and the indirect methods. The black curve represents the workload process estimated after observing the data gathered throughout the day. The blue and red curves represents the workload forecasts for the indirect and direct forecasts, respectively, given data up to 12 PM.*

variance does not change drastically (see Figure 1).

5.7. *Confidence bands.*   Following Weinberg, Brown and Stroud (2007), we introduce the 95% confidence band coverage (COVER) and the average 95% confidence band width (WIDTH). Specifically, for each day j, let

$$COVER_j = \frac{1}{K} \sum_{k=1}^{K} I\left(F_{L,jk} < N_{jk} < F_{U,jk}\right) \; ; \; WIDTH_j = \frac{1}{K} \sum_{k=1}^{K} \left(F_{U,jk} - F_{L,jk}\right) \; ,$$

Example 3	RMSE			APE		
	Day ahead	10 AM	12 PM	Day ahead	10 AM	12 PM
Minimum	3.66	3.62	3.92	4.47	4.33	4.60
Q1	5.37	5.63	5.05	5.57	5.41	5.64
Median	6.80	7.01	6.87	6.71	6.84	6.31
Mean	7.64	7.19	6.97	7.23	7.10	6.97
Q3	9.01	8.97	8.59	8.83	8.16	7.44
Maximum	16.12	11.84	11.13	12.17	11.80	12.46

TABLE 3. *Summary of statistics (minimum, lower quartile (Q1), median, mean, upper quartile (Q3), maximum) of RMSE and APE for the forecast based on the mean of the previous days and the BLUP, using 10 AM and 12 PM cuts for the weekends data set.*

where $(F_{L,jk}, F_{U,jk})$ is the confidence band of day j, evaluated at the beginning of the k-th interval (see (16)). The mean coverage and mean width, for all three examples, are presented in Table 4. First, note that for all three examples, the width of the confidence band narrows down as more information is revealed. In other words, the width of the confidence band for the 12 PM forecast is narrower than the width for the 10 AM forecast which, in turn, is narrower than the width for the pervious days' mean. We also see that the mean coverage becomes more accurate as more information is revealed. Figure 4 depicts the confidence bands for the arrival process on a particular Sunday. Note that the confidence bands for the previous days' forecast and the 10 AM forecast almost coincide. However, at 12 PM, when enough information on this particular day becomes available, the confidence band narrows down and does capture the underlying behavior.
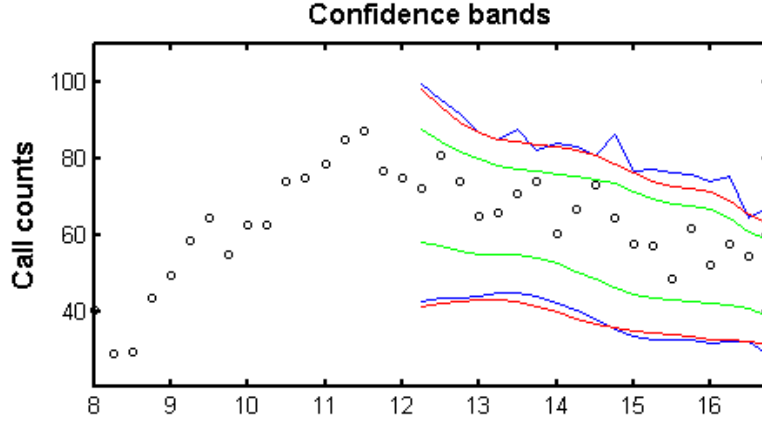

FIG 4. *Confidence bands for Sunday, August 10, 2003. The black dots represent the true call counts in fifteen-minutes resolution. The confidence bands based on previous days, 10 AM data, and 12 PM data are represented by the blue, red, and green lines, respectively.*

	Coverage			Width		
	Example 1	Example 2	Example 3	Example 1	Example 2	Example 3
Mean	93.19%	91.69%	98.15%	79.73	62.80	40.15
10 AM	94.14%	92.27%	98.64%	74.99	56.45	39.53
12 PM	94.86%	93.08%	96.49%	73.07	55.95	31.40

TABLE 4. *The mean confidence band coverage and the mean width for the forecasts based on the previous days' mean, the 10 AM cut and the 12 PM cut for the arrival process on the working days data set (Example 1), the workload process on the working days data set (Example 2) and the arrival process on the weekends data set (Example 3).*

Summarizing, using call center data, we demonstrated that forecasting of curve continuation can be achieved successfully by the proposed BLUP. We also showed that confidence bands for such forecasts can be obtained using cross-validation.

6. Concluding Remarks. We have constructed the best linear unbiased predictor (BLUP) for the continuation of a curve. We now add the following comments regarding the construction of the BLUP and its application to call center data.

First, in our analysis we have used a spline model to describe the functions. This is not required for the construction of the BLUP, and the proof of Theorem 3.4 holds for any function space of finite dimension. However, as discussed previously, there are two main advantages of using spline representation. First, the computation of the covariance operators, for S_1, S_2 and S and between them, as well as the pseudo-inverse covariance operator Γ_{11}^+, are all computationally simple when working with splines. Second, the representation of the restriction of a function to a partial segment does not suffer from collinearity of the basis functions, which can be the case for a more general setting. Indeed, the number of basis elements can be reduced significantly in the spline function model, depending on the number of knots in the partial segment, while the number of basis elements could remain the same in a more general model.

Second, we have assumed that the random function X lies within a function space of (possibly large) finite-dimension. While this is a restrictive assumption, there are some difficulties with the BLUP definition (and computation) for a random function that lies in an infinite-dimensional space. The main difficulty is that inverting the covariance operator (as done in Lemma 3.3 for finite dimension) is problematic since the inverse of the covariance operator need not be bounded and may not exists. However, we believe that characterization of the BLUP in the infinite-dimension case is possible under some conditions. Further research is required to address this question.

Finally, in this work we forecasted the continuation of the workload process. As discussed in Feldman *et al.* (2008) and Reich (2010), the workload process is a more appropriate candidate than the arrival process, as a basis for determining staffing levels in call centers. This work, along with Aldor-Noiman, Feigin and Mandelbaum (2009) and Reich (2010), are the first steps in exploring direct forecasting of the workload process, but more remains to done (see, for example Whitt, 1999; Zeltyn *et al.*, 2009).

APPENDIX A: PROOFS

A.1. Lemma A.1.

LEMMA A.1. *Let T be a $n \times p$ matrix of rank s and let L be a $p \times p$ positive definite diagonal matrix. Then the following assertions are true*

1. $T'T(T'T)^+ T' = T'$
2. $T'LT(T'LT)^+ T' = T'$
3. $(T'T)^+ = T' \left((TT')^+\right)^2 T$

PROOF.   1. If $T'T$ is invertible then $(T'T)^+ = (T'T)^{-1}$ and the result follows. Otherwise, let $U\Lambda V'$ be the svd (singular value decomposition, see Golub and Loan, 1983) of T where U and V are orthonormal columns matrices of size $n \times s$ and $p \times s$ respectively, and Λ is a $s \times s$ positive definite diagonal matrix. Then

$$
\begin{aligned}
T'T(T'T)^+ T' &= (V\Lambda U')(U\Lambda V')\left((V\Lambda U')(U\Lambda V')\right)^+ V\Lambda U' \\
&= V\Lambda^2 V'(V\Lambda^2 V')^+ V\Lambda U' .
\end{aligned}
$$

Since Λ is invertible and V has orthonormal columns $(V\Lambda^2 V')^+ = V\Lambda^{-2}V'$. Hence

$$
T'T(T'T)^+ T' = V\Lambda^2 V'V\Lambda^{-2}V'V\Lambda U' = V\Lambda U' = T' .
$$

2. Denote $W = L^{1/2}T$, then $T'LT(T'LT)^+ T' = W'W(W'W)^+ W'L^{-1/2}$. Using 1., we obtain $W'W(W'W)^+ W'L^{-1/2} = W'L^{-1/2} = T'$.

3. Since TT' is a positive semi-definite matrix, TT' has a spectral representation of the form $TT' = \sum_{i=1}^s \lambda_i v_i v_i'$ where $s \leq \min\{n,p\}$, $\lambda_i > 0$ and $\{v_i\}$ is an orthonormal set of vectors. Note that $TT'v_i = \lambda_i v_i$ and hence $T'T(T'v_i) = \lambda_i T'v_i$. Moreover $\|T'v_i\|^2 = v_i'TT'v_i = v_i'(\lambda_i v_i) = \lambda_i$. Hence, we obtained that $\{T'v_i/\sqrt{\lambda_i}\}$ is the set of orthonormal eigenvectors of $T'T$ with the respective non-zero eigenvalues $\{\lambda_i\}$. Thus,

$$
T'T = \sum_{i=1}^s \lambda_i \frac{T'v_i}{\sqrt{\lambda_i}} \left(\frac{T'v_i}{\sqrt{\lambda_i}}\right)' = T' \left(\sum_{i=1}^s v_i v_i'\right) T .
$$

Using the spectral representation we also have

$$
T' \left((TT')^+\right)^2 T = T' \left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right) T .
$$

In order to show that $(T'T)^+ = T' \left((TT')^+\right)^2 T$ we need to show the following (see Golub and Loan, 1983):

(a) $(T'T)\left(T'\left((TT')^+\right)^2 T\right)(T'T) = (T'T)$

(b) $\left(T'\left((TT')^+\right)^2 T\right)(T'T)(T'\left((TT')^+\right)^2 T) = \left(T'\left((TT')^+\right)^2 T\right)$

(c) $\left((T'T)\left(T'\left((TT')^+\right)^2 T\right)\right)' = (T'T)(T'\left((TT')^+\right)^2 T)$

(d) $\left(\left(T'\left((TT')^+\right)^2 T\right)(T'T)\right)' = (T'\left((TT')^+\right)^2 T)(T'T)$

In order to see (a), note that

$$
\begin{aligned}
(T'T)\left(T'\left((TT')^+\right)^2 T\right)(T'T) &= T'(TT')\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)(TT')T \\
&= T'\left(\sum_{i=1}^s \lambda_i v_i v_i'\right)\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)\left(\sum_{i=1}^s \lambda_i v_i v_i'\right)T \\
&= T'\left(\sum_{i=1}^s v_i v_i'\right)T = T'T.
\end{aligned}
$$

Similarly, for (b), we have

$$
\begin{aligned}
\left(T'\left((TT')^+\right)^2 T\right)(T'T)\left(T'\left((TT')^+\right)^2 T\right) &= \\
&= T'\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)(TT')^2\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)T \\
&= T'\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)\left(\sum_{i=1}^s \lambda_i v_i v_i'\right)^2\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)T \\
&= T'\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)T = T'\left((TT')^+\right)^2 T.
\end{aligned}
$$

For (c),

$$
\begin{aligned}
\left((T'T)\left(T'\left((TT')^+\right)^2 T\right)\right)' &= \left(T'(TT')\left(\sum_{i=1}^s \lambda_i^{-2} v_i v_i'\right)T\right)' = \left(T'\left(\sum_{i=1}^s \lambda_i^{-1} v_i v_i'\right)T\right)' \\
&= T'\left(\sum_{i=1}^s \lambda_i^{-1} v_i v_i'\right)T = (T'T)\left(T'\left((TT')^+\right)^2 T\right).
\end{aligned}
$$

Finally, (d) is shown similarly to (c).

$\square$

A.2. Proof of Lemma 3.3.

PROOF. By (2) we may write $Y_1(t) - \mu(t) = \boldsymbol{b_1}(t)'(A_1\boldsymbol{h} + B_1\boldsymbol{\epsilon})$. Hence,

$$
\left(\Gamma_{11}\Gamma_{11}^+ \ (A_1\boldsymbol{h} + B_1\boldsymbol{\epsilon}) \right)(t) =
$$

$$
= \boldsymbol{b_1}(t)'G_{11}W_1W_1^{-1}G_{11}^+(A_1\boldsymbol{h} + B_1\boldsymbol{\epsilon})
$$

$$
= \boldsymbol{b_1}(t)'G_{11}G_{11}^+(A_1\boldsymbol{h} + B_1\boldsymbol{\epsilon})
$$

$$
= \boldsymbol{b_1}(t)'[A_1, B_1] \begin{bmatrix} L & 0 \\ 0 & \Sigma \end{bmatrix} \begin{bmatrix} A_1' \\ B_1' \end{bmatrix} \left([A_1, B_1] \begin{bmatrix} L & 0 \\ 0 & \Sigma \end{bmatrix} \begin{bmatrix} A_1' \\ B_1' \end{bmatrix} \right)^+ \begin{bmatrix} A_1\boldsymbol{h} \\ B_1\boldsymbol{\epsilon} \end{bmatrix}
$$

and the result follows from Lemma A.1.

Substituting $\boldsymbol{h} = LA_2'$ and $\boldsymbol{\epsilon} = 0$ in the last equation, we also obtain

$$
(17) \qquad\qquad G_{11}G_{11}^+g_{12} = \Gamma_{11}G_{11}^+(A_1LA_2' + B_1\boldsymbol{0}) = g_{12}\,.
$$

$\square$

ACKNOWLEDGEMENTS

We thank Michael Reich for helpful discussions and for providing us with the data for the workload example.

SUPPLEMENTARY MATERIAL

Supplement A: Code and data sets

(http://pluto.huji.ac.il/ yaacov/blup.zip). Please read the file README.pdf for details on the files in this folder.

REFERENCES

ALDOR-NOIMAN, S., FEIGIN, P. D. and MANDELBAUM, A. (2009). Workload forecasting for a call center: Methodology and a case study. To appear.

ANTONIADIS, A., PAPARODITIS, E. and SAPATINAS, T. (2006). A functional waveletkernel approach for time series prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 837–857.

BESSE, P., CARDOT, H. and FERRATY, F. (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Computational Statistics & Data Analysis* **24** 255–270.

BESSE, P. C., CARDOT, H. and STEPHENSON, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* **27** 673–687.

DE BOOR, C. (2001). *A practical guide to splines*, Revised ed. *Applied Mathematical Sciences*. Springer-Verlag New York.

DONIN, O., FEIGIN, P. D., MANDELBAUM, A., ZELTYN, S., TROFIMOV, V., ISHAY, E., KHUDIAKOV, P. and NADJHAROV, E. (2006). The Call Center of US Bank. Avaliable at http://ie.technion.ac.il/Labs/Serveng/files/The_Call_Center_of_US_Bank.pdf.

FELDMAN, Z., MANDELBAUM, A., MASSEY, W. A. and WHITT, W. (2008). Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science* **54** 324–338.

GANS, N., KOOLE, G. and MANDELBAUM, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing Service Operations Management* **5** 79-141.

GOLUB, G. H. and LOAN, C. F. V. (1983). *Matrix computations*. Johns Hopkins University Press, Baltimore, Maryland.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

KNAFL, G., SACKS, J. and YLVISAKER, D. (1985). Confidence bands for regression functions. *Journal of the American Statistical Association* **80** 683–691.

KNEIP, A. (1994). Nonparametric estimation of common regressors for similar curve data. *The Annals of Statistics* **22** 1386–1427.

MARSAGLIA, G. (1964). Conditional means and covariances of normal variables with singular covariance matrix. *Journal of the American Statistical Association* **59** 1203–1204.

RAMSAY, J. and SILVERMAN, B. W. (2002). *Applied functional data analysis: methods and case studies*, 2nd ed. *Springer Series in Statistics*. Springer-Verlag New York.

RAMSAY, J. and SILVERMAN, B. W. (2005). *Functional data analysis*. *Springer Series in Statistics*. Springer-Verlag New York.

REICH, M. (2010). The workload process: modelling, inference and applications Master's thesis, Technion - Israel Institute of Technology. In preparation. The proposal is avaliable at http://ie.technion.ac.il/serveng/References/references.html.

ROBINSON, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6** 15–32.

ROZENSHMIDT, L. (2008). On priority queues with impatient customers: Stationary and time-varying analysis Master's thesis, Technion - Israel Institute of Technology. Available at http://iew3.technion.ac.il/serveng/References/thesis_Luba_Eng.pdf.

SANSONE, G. (1991). *Orthogonal functions*, Rev. ed. ed. Dover Publications,, New York.

SHEN, H. (2009). On modeling and forecasting time series of smooth curves. *Technometrics* **51** 227–238.

SHEN, H. and HUANG, J. Z. (2008). Interday Forecasting and Intraday Updating of Call Center Arrivals. *Manufacturing Service Operations Management* **10** 391–410.

WEINBERG, J., BROWN, L. D. and STROUD, J. R. (2007). Bayesian forecasting of an inhomogeneous poissonprocess with applications to call center data. *Journal of the American Statistical Association* **Vol. 102**.

WHITT, W. (1999). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* **24** 205 - 212.

ZELTYN, S. (2005). Call centers with impatient customers: Exact analysis and many-server asymptotics of the M/M/n+G queue. PhD thesis, TechnionIsrael Institute of Technology. Available at http://ie.technion.ac.il/serveng/References/references.html.

ZELTYN, S., CARMELI, B., GREENSHPAN, O., MESIKA, Y., WASSERKRUG, S., VORTMAN, P., MARMOR, Y. N., MANDELBAUM, A., SHTUB, A., LAUTERMAN, T., SCHWARTZ, D., MOSKOVITCH, K., TZAFRIR, S. and BASIS, F. (2009). Simulation-Based Models of Emergency Departments: Operational, Tactical and Strategic Staffing. Under review.

Yair Goldberg and Ya'acov Ritov
Department of Statistics and
The Center for the Study of Rationality
The Hebrew University
Jerusalem 91905, Israel
E-mail: yair.goldberg@mail.huji.ac.il
        yaacov.ritov@huji.ac.il

Avishai Mandelbaum
Industrial Engineering and Management
Technion-Israel Institute of Technology
Haifa 32000, Israel
E-mail: avim@tx.technion.ac.il