Bias-corrected Estimation of the Density of a Conditional Expectation in Nested Simulation Problems

RAN YANG, Department of Industrial Engineering & Management Sciences, Northwestern University
DAVID KENT, Department of Statistics and Data Science, Cornell University
DANIEL W. APLEY and JEREMY STAUM, Department of Industrial Engineering & Management Sciences, Northwestern University
DAVID RUPPERT, Department of Statistics and Data Science and School of Operations Research and Information Engineering, Cornell University

Many two-level nested simulation applications involve the conditional expectation of some response variable, where the expected response is the quantity of interest, and the expectation is with respect to the inner-level random variables, conditioned on the outer-level random variables. The latter typically represent random risk factors, and risk can be quantified by estimating the probability density function (pdf) or cumulative distribution function (cdf) of the conditional expectation. Much prior work has considered a naïve estimator that uses the empirical distribution of the sample averages across the inner-level replicates. This results in a biased estimator, because the distribution of the sample averages is over-dispersed relative to the distribution of the conditional expectation when the number of inner-level replicates is finite. Whereas most prior work has focused on allocating the numbers of outer- and inner-level replicates to balance the bias/variance trade-off, we develop a bias-corrected pdf estimator. Our approach is based on the concept of density deconvolution, which is widely used to estimate densities with noisy observations but has not previously been considered for nested simulation problems. For a fixed computational budget, the bias-corrected deconvolution estimator allows more outer-level and fewer inner-level replicates to be used, which substantially improves the efficiency of the nested simulation.

CCS Concepts: • **Computing methodologies → Simulation theory**; • **Applied computing → Forecasting**; • **Networks → Data path algorithms;**

Additional Key Words and Phrases: Deconvolution, estimating a conditional variance, quadratic programming, shape constraints, Stein's unbiased risk estimation, rate of convergence

ACM Reference format:
Ran Yang, David Kent, Daniel W. Apley, Jeremy Staum, and David Ruppert. 2021. Bias-corrected Estimation of the Density of a Conditional Expectation in Nested Simulation Problems. *ACM Trans. Model. Comput. Simul.* 31, 4, Article 22 (July 2021), 36 pages.
https://doi.org/10.1145/3462201

The authors gratefully acknowledge the support from NSF grant AST-1814840.
Authors' addresses: R. Yang, D. W. Apley (corresponding author), and J. Staum, Department of Industrial Engineering & Management Sciences, Northwestern University, 2145 Sheridan Road, Evanston, IL, 60208; emails: ranyang2011@u.northwestern.edu, {apley, j-staum}@northwestern.edu; D. Kent, Department of Statistics and Data Science, Cornell University, Ithaca, NY, 14853; email: dk657@cornell.edu; D. Ruppert, Department of Statistics and Data Science and School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, 14853; email: dr24@cornell.edu.

1049-3301/2021/07-ART22 $15.00
https://doi.org/10.1145/3462201

1 INTRODUCTION

Nested Monte Carlo simulations (a.k.a. two-level simulations) are widely used for risk assessment in engineering, finance, and other areas. To make things more concrete, consider the example of portfolio risk assessment, in which the outer level draws random realizations (scenarios i) of histories of cash flows, commodity, and equity prices, and other random risk factors for each position in the portfolio, from current time up to some specified future time horizon. The intent is to estimate the future value of the portfolio at the horizon for each specific risk factor realization. To estimate the portfolio value at the horizon for one risk factor realization (which corresponds to one outer-level replicate), an inner-level simulation draws multiple random realizations (a different realization j for each inner-level replicate) of position behavior beyond the horizon up to some final maturity time. Variation across the outer simulations is financial risk and is of interest. Variation among inner-level simulations is Monte Carlo measurement error and, as will be explained, causes bias that needs to be corrected.

Let n denote the number of outer-level replicates, m_i the number of inner-level replicates on the ith outer-level replicate (which we refer to as the ith scenario), and Y_{ij} the scalar simulated output response of interest for the jth inner-level replicate of the ith scenario.

In the example, the response Y_{ij} would be the simulated value of all positions in the portfolio at maturity time for that inner-level realization j within the ith outer-level simulation, plus various things like accrued cash flows at the horizon, and then discounted back to the current time. Conditioned on the outer-level random quantities for scenario i, the repricing of positions for scenario i that will occur at the horizon is based on their simulated values $\{Y_{ij} : j = 1, 2, \ldots, m_i\}$ at the maturity time. Specifically, if we define X_i as the conditional expectation of Y_{ij} across a hypothetically infinite number of inner-level replicates, conditioned on scenario i, then this conditional expectation determines the portfolio value (and thus the profit or loss) at the horizon for the scenario i risk factor realizations. Consequently, the portfolio value X_i for scenario i at the horizon is a conditional expectation, and the pdf f_X (with respect to the outer-level random quantities, i.e., across i) of this conditional expectation is used to compute all risk-related quantities, such as the expected loss, probability of large losses or gains, loss quantiles, and so on. Note that in reality, when the horizon time arrives, the portfolio would be priced by conducting a single-level simulation of what happens from the horizon to the final maturity time, with the initial state at the horizon being whatever state reality was in at that horizon time. For this single-level simulation, one could use a large number of replicates to ensure that the sample average of $\{Y_{i,j} : j = 1, 2, \ldots\}$ is sufficiently close to its conditional expectation, given the scenario i that represents the realization of real-life events up to the horizon time. In contrast, since we must use a two-level simulation conducted at the current time to model how the portfolio might be priced at the future horizon (which is dictated by the outer-level scenario), we cannot afford to conduct a large number of inner-level replicates for each of a large number of scenarios. We elaborate on this example below and in Section 3.1.1. Also see References [12, 20, 21] for much more detailed descriptions of estimating the distribution of the discounted value of portfolio profit and loss at some future time horizon and the nested simulation framework.

The general two-level simulation setting that we consider in this work is broadly applicable anytime one uses a standard (single-level) Monte Carlo simulation to estimate an expectation via a Monte Carlo average, where the expectation depends on other exogenous phenomena that are random or uncertain. The inner level becomes the Monte Carlo simulation of interest, and each outer-level replicate draws a different realization of the exogenous phenomena from their assumed distribution. See Reference [4] for applications in healthcare. In Section 3.1.2, we discuss queuing systems applications in which the outer-level random quantities are uncertain parameters of the distributions of simulation inputs.

Our primary goal for the general setting is to estimate the pdf f_X of the conditional expectation X (we drop the outer-level replicate subscript i and assume outer-level replicates are i.i.d.), based on the response observations $\{Y_{ij} : i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m_i\}$ from the nested simulation. Note that the estimated pdf can be integrated numerically to estimate the cdf $F_X(x) = \Pr[X < x] = \int_{-\infty}^{x} f_X(u)\mathrm{d}u$, which in turn can be inverted numerically to estimate quantiles. Since the average response

$$\bar{Y}_i \equiv \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij} \tag{1}$$

for the ith scenario can be viewed as an estimate of X_i, a commonly used naïve estimator of f_X is the empirical density of $\{\bar{Y}_i : i = 1, 2, \ldots, n\}$, e.g., its histogram or some kernel density estimator.

The obvious shortcoming of this naïve estimator, and the primary motivation for this work, is that $\bar{Y}_i$ is only a noisy estimator of X_i for finite m_i, in which case the empirical distribution of $\{\bar{Y}_i : i = 1, 2, \ldots, n\}$ will be over-dispersed (i.e., have a larger spread, larger variance, etc.), relative to f_X. This translates to a *bias* in the estimator of $f_X(x)$, because at the tail x values the estimator is consistently higher than the true $f_X(x)$. The over-dispersion bias is more severe with smaller m_i, which is problematic, because it may be computationally prohibitive to use a large m_i when n must also be sufficiently large. We illustrate the over-dispersion bias in Figure 1 for the portfolio risk example (the details of the setting and how $Y_{i,j}$ are defined and computed are discussed in Section 3.1.1). Figure 1 shows the true f_X for this example, along with a histogram of $\{\bar{Y}_i : i = 1, 2, \ldots, n\}$ and the corresponding naïve estimator (a kernel density estimator). Note that the naïve estimator differs substantially from f_X and is much more dispersed. In terms of its impact on risk assessment, if X is a portfolio loss, then the naïve estimator would overestimate the probability of large and small losses and underestimate the probability of moderate losses. As a preview, we also show the estimator that we develop in this article, which is a type of deconvolution estimator intended to correct for the bias present in the naïve estimator. From Figure 1, our bias-corrected deconvolution estimator is far closer to the true f_X.

Some prior works have investigated how to most efficiently estimate characteristics of f_X with a limited computational budget, since it is computationally expensive to have both large n and large m_i. Most such prior work has used the biased naïve estimator and focused on how to balance between outer-level (n) and inner-level (m_i) replicates. A standard approach is to minimize the **mean squared error (MSE)** of the naïve estimator of some specific functional of f_X by balancing the tradeoff between large bias (worse for small m_i) versus large variance (usually worse for small n). References [13, 20, 21] focused on situations where the inner-level sample size was constant across all outer scenarios, which they referred to as uniform allocation. They showed that in a certain asymptotic sense, the bias and variance of the naïve estimator are functions of m and n, respectively. For a fixed computational budget, they derived the asymptotically optimal allocation between n and m to minimize the overall MSE of a specific characteristic of the naïve estimator (e.g., the probability that X exceeds some specified value). There are also works (e.g., References [13, 18, 21]) studying naïve estimators in the nonuniform allocation situation in which the number of inner-level replicates varies across scenarios and is allocated adaptively. When the goal is to estimate a tail probability or quantile of f_X, References [13, 18, 21] showed that nested simulations can be more efficient by allocating more computational budget to scenarios for which $\bar{Y}$ falls near the tail. Reference [5] further considered the naïve estimator in the nonuniform/adaptive allocation situation. To estimate $\Pr[X > c]$ for some specified c, they developed an adaptive allocation procedure that uses larger m_i on scenarios for which $\bar{Y}_i$ is closer to c and/or the uncertainty in $\bar{Y}_i$ is larger, and they demonstrated a lower asymptotic bias and MSE of the naïve estimator for this nonuniform allocation.

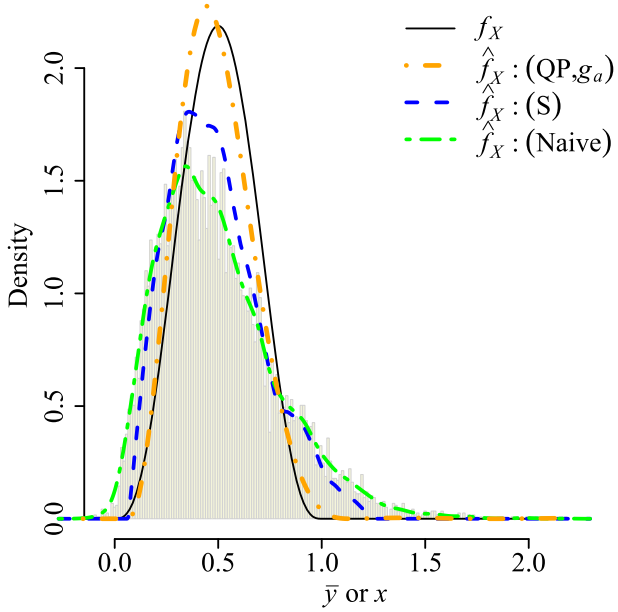


Fig. 1. For the portfolio loss example with $m_i = 5$ and $n = 10^4$, illustration of over-dispersion bias for the naïve estimator (green double-dashed curve labeled Naive), which is a smoothed version of the histogram shown for $\bar{Y}_i$ $(i = 1, \ldots, n)$ and is much more dispersed than the true pdf f_X (solid black curve), because $\bar{Y}$ is a noisy version of X. Our deconvolution-based estimator (orange dot-dashed curve labeled QP,ga) corrects for this bias and is much closer to the true f_X. The blue dashed curve (labeled S) is the estimator of Reference [28].

Two aspects of the approach of Reference [5], upon which we seek to improve, are that it relies on the biased naïve estimator and that it optimizes the allocation for one specific feature of f_X ($\Pr[X > c]$ for one specific c). Regarding the latter, multiple characteristics of f_X or even the entire f_X may be of interest, in which case an allocation that increases m_i when $\bar{Y}_i$ is close to one specific value c will not be appropriate. Regarding the former, if one could modify the estimator to remove the bias, then it is reasonable to suppose that a lower MSE could be achieved for the same computational expense. In essence, with an unbiased estimator, it is not necessary to trade a larger variance for a smaller bias. Under a uniform allocation, Reference [30] developed an unbiased estimator of the variance of X. They showed that with the unbiased estimator, minimizing the MSE under a fixed computational budget $M = nm$ generally resulted in a much smaller optimal m and larger optimal n than if the naïve biased estimator is used. In fact, they showed that even as $M \to \infty$, the optimal m remains bounded and surprisingly small. This finding allows a much larger n to be used in the nested simulation, which ultimately gives a much better estimator with smaller $\mathrm{Var}(X)$ for the same computational budget. Whereas Reference [30] considered only $\mathrm{Var}(X)$ and its unbiased estimator, in this article, we consider the much harder problem of correcting the bias in the estimation of the entire pdf f_X. This is important, because the entire distribution of X, or at least multiple characteristics of it, are typically of interest. We refer to this as bias-corrected density estimation, because it reduces or removes the over-dispersion bias when using the noisy observations $\{\bar{y}_i : i = 1, 2, \ldots, n\}$ to estimate f_X.

Importantly, we show that the conclusions reached by Reference [30] regarding using two-level simulation to estimate $\mathrm{Var}(X)$ apply to our estimation of f_X as well. Namely, under uniform allocation, quite reasonable estimation of f_X, F_X, and quantiles of X can be achieved with surprisingly

small m; and for a fixed computational budget, more of the budget should be allocated to increasing the number of outer-level scenarios (larger n) and less to inner-level replicates (smaller m). This is relative to the recommendations from the prior work that have used the biased estimator (e.g., References [5, 13]), for which the optimal m continues to grow as the computational budget grows. By allowing more outer-level replicates for a given computational budget, our approach reduces the variance of the estimator without substantially increasing the bias, since the bias is corrected.

Some prior work has considered in a limited manner the bias in the estimation of f_X in the context of nested simulation. Reference [13] developed an expression for the bias in the naïve estimator of $F_X(\cdot)$, and Reference [5] used the same bias expression when analyzing the effects of n and m on the MSE. However, neither of these works used the expression to develop a better estimator by correcting for the bias in the naïve estimator. The bias is "notoriously difficult to estimate" according to Reference [5], and presumably this is the reason that they did not incorporate it as the basis for a modified estimator with reduced bias. In contrast, we show in this article that the structure of the bias can be represented quite naturally within the framework of density deconvolution, and the resulting bias-corrected deconvolution estimator performs quite reliably and much better than the naïve estimator. Reference [13] did consider a modified naïve estimator of F_X based on a jackknife approach and showed that it can reduce the bias in the estimator, at the expense of only a modest increase in the variance. Reference [28] constructed a bias-corrected estimator of f_X based on a type of kernel smoothing density estimator. They showed that their bias-corrected estimator is less computationally expensive than a typical jackknife estimator and proved it to have a better MSE convergence rate than the naïve kernel smoothing estimator, given their optimal m and n allocation.

In this article, we demonstrate that our deconvolution-based estimator outperforms existing methods for estimating f_X in the two-level simulation setting. An additional advantage of our deconvolution estimator is that it is straightforward to incorporate a number of common density constraints such as nonnegativity, integration-to-one, unimodality, tail convexity, tail monotonicity, and support constraints (e.g., that $X \geq 0$), which can further substantially improve the density estimation. In addition to focusing on a bias-corrected estimator, the scope of this work also differs from that of References [5, 13] in that we estimate the entire f_X, as opposed to optimizing the choice of n and m_i for one specific feature of f_X. Practitioners are often interested in many different features of the distribution of X, such as the mean, variance, probability of large loss (or large gain), $F_X(c)$ for a number of different c, value-at-risk for many different α levels, expected shortfall for many different α levels, and so on. To calculate many different features, and more generally to simply understand the distribution of X, having a good estimator of f_X that is unbiased even for small m is much more useful than having an estimator of only one specific feature of f_X.

Our main contributions are to recast the problem of estimating the distribution of the conditional expectation in two-level simulations as a density deconvolution problem to adapt existing deconvolution estimators to this nested simulation problem and to demonstrate its usage and its performance for this purpose. In spite of the excellent performance that we demonstrate for this problem, the connection between deconvolution and the density of a conditional expectation in two-level simulations has evidently not been recognized and developed before. This is perhaps because it is not quite an off-the-shelf application of deconvolution, for reasons that become evident in Sections 2.1 and 2.3. The remainder of the article is organized as follows: Section 2 introduces the deconvolution framework for the nested simulation problem and discusses various considerations in developing the deconvolution estimator for this problem. In Section 3, we illustrate the performance of the estimator and compare it with the naïve estimator and the estimator of Reference [28] under both uniform and adaptive allocations of n and m_i. Some concluding remarks are presented in Section 4.

2 DECONVOLUTION MODELING FRAMEWORK FOR NESTED SIMULATION

We cast the estimation of f_X from the noisy observations $\{\bar{Y}_i : i = 1, 2, \ldots, n\}$ as a density deconvolution problem, for which a number of methods (e.g., References [8, 29, 36]) have been developed. Although certain Fourier or kernel deconvolution based methods have desirable theoretical properties like asymptotic convergence rates, they sometimes do not perform well in practice. Using a variety of examples, Reference [36] demonstrated numerically that their **quadratic programming (QP)**-based deconvolution estimator generally performed much better for the finite samples that one analyzes in practice. In this article, we focus on the QP deconvolution estimator of Reference [36] because of its superior performance and also because it allows a number of common shape constraints on the distribution to be incorporated to further improve the performance of the estimator and the efficiency of the nested simulation.

In this section, we discuss how the nested simulation problem can be cast as a deconvolution problem and how the QP deconvolution method of Reference [36] can be adapted to solve it. In Section 2.1 the relationship between deconvolution and density estimation in the nested simulation framework is established. In Section 2.2, we derive the QP deconvolution estimator for the density of a conditional expectation in a nested simulation, and in Section 2.3, we discuss how to estimate the conditional error variance function.

2.1 Casting Nested Simulation as a Convolution/Deconvolution Problem

To facilitate the development, we introduce the following notation and write,

$$Y_{ij} = Y(\omega_i, \xi_{i,j}), \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m_i, \tag{2}$$

where ω_i denotes the random outcome that determines ith scenario, and $\xi_{i,j}$ denotes the further random outcome, which, together with ω_i, determines the response Y_{ij} for the jth inner-level replicate of the ith scenario. In the portfolio risk example, the random outcome ω_i represents the complete set of conditions that determine the entire realization of risk factors for scenario i (e.g., complete price history from current time to the horizon of all commodities, equities, etc., in the portfolio, together with factors like interest rates); and the random outcome $\xi_{i,j}$ represents the same but from the horizon to the final maturity time, for inner-level replicate j of scenario i. We assume $\{\omega_i : i = 1, 2, \ldots, n\}$ are independent, identically distributed (i.i.d.), and given ω_i, $\{\xi_{i,j} : j = 1, 2, \ldots, m_i\}$ are conditionally i.i.d. for each i. Further assume that for $i \neq k$, $\xi_{i,j}$ and $\xi_{k,l}$ are independent for all j and l. The number of inner-level replicates m_i may depend on ω_i. This setting is quite general in that the outer-level random variables and events determined by ω may be parametric with few or many parameters, nonparametric, infinite dimensional, and so on. In Section 3.1.2, we consider a parametric special case in which ω represents uncertainty in the parameters of an input distribution in a stochastic call center simulation model. Note that, after each scenario i, the corresponding ω_i could be observed in the simulation if desired, since ω_i represents all of the underlying random variables that determine the simulation results for scenario i. However, our approach does not use observations of ω_i directly. We only use the observed $\{Y_{i,j} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, m_i\}$.

The goal is to estimate the distribution (with respect to ω) of the conditional expectation $X = X(\omega) = \mathrm{E}[Y|\omega]$, based on observations of $\{Y_{ij} : i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m_i\}$. Because of the i.i.d. assumption, we have dropped the subscripts indicating the inner-level and outer-level replicates in $X = \mathrm{E}[Y|\omega]$, which we will do throughout the article unless subscripts are needed to distinguish the replicate. To develop the convolution relationship between the pdfs of $\bar{Y}$ and X, which are at the heart of the density deconvolution approach, we write

$$
\begin{aligned}
Y_{ij} &= X_i + \epsilon_{ij} \quad \text{and} && (3)\\
\bar{Y}_i &= X_i + Z_i, \quad \text{where} && (4)\\
X_i &= X(\omega_i) \equiv \mathrm{E}[Y_{ij}|\omega_i], && (5)\\
\epsilon_{ij} &= \epsilon(\omega_i, \xi_{i,j}) \equiv Y_{ij} - \mathrm{E}[Y_{ij}|\omega_i], && (6)\\
Z_i &= Z(\omega_i, \xi_{i,1}, \xi_{i,2}, \ldots, \xi_{i,m_i}) \equiv \bar{Y}_i - \mathrm{E}[Y_{ij}|\omega_i] = \frac{1}{m_i}\sum_{j=1}^{m_i} \epsilon_{ij}. && (7)
\end{aligned}
$$

The pdf of $\bar{Y}$ is related to the pdfs of X and Z via

$$f_{\bar{Y}}(y) = \int_{-\infty}^{\infty} f_{\bar{Y}|X}(y|x) f_X(x)\mathrm{d}x = \int_{-\infty}^{\infty} f_{Z|X}(y-x|x) f_X(x)\mathrm{d}x, \tag{8}$$

where the subscripts on the pdfs indicate marginal, conditional, or joint densities. The deconvolution approach described in the next section uses Equation (8) to obtain an estimate of f_X from an estimate of $f_{\bar{Y}}$, the latter being available from the $\{\bar{Y}_i : i = 1, 2, \ldots, n\}$ observations. Relating the unknown f_X to $f_{\bar{Y}}$ in Equation (8) requires the conditional distribution $f_{Z|X}$. Although this is unknown in general, in our nested simulation setting, we can use the following approximation. Because $\bar{Y}|\omega$ is a sample average in our setting, the central limit theorem suggests that we can approximate $f_{Z|\omega}$ as a normal distribution with mean zero and variance $\sigma_Z^2(\omega) \equiv \mathrm{Var}[Z|\omega] = \mathrm{Var}[\bar{Y}|\omega]$. If each $f_{Z|\omega}$ is normal, then the distribution $f_{Z|X}$ needed in Equation (8) is then the mixture of zero-mean normal distributions

$$f_{Z|X}(z|x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_Z(\omega)}\phi\left(\frac{z}{\sigma_Z(\omega)}\right)\mathrm{d}P_{\omega|X}(\omega|x), \tag{9}$$

where $\phi(\cdot)$ denotes the standard normal density, and $P_{\omega|X}(\omega|x)$ denotes the conditional distribution of ω, given that $X(\omega) = x$. Although this mixture of zero-mean normal distributions is not exactly normal, it should typically be close to normal. In particular, in simulation settings in which the conditional variance of $\bar{Y}|\omega$ depends on ω only via the conditional mean $X(\omega) = \mathrm{E}[\bar{Y}|\omega]$, the mixture of normals in Equation (9) is exactly normal with mean zero and with variance $v(x) \equiv \mathrm{Var}[Z|X = x]$, where $v(x) = \sigma_Z^2(\omega)$ for any ω such that $X(\omega) = x$. Consequently, if the conditional variance of $\bar{Y}|\omega$ depends on ω predominantly as a function of the conditional mean (e.g., if the conditional standard deviation of $Y|\omega$ is approximately proportional to its mean, which is a common form of heteroskedasticity), then a normal approximation for $f_{Z|X}$ should be reasonable.

The following arguments provide further justification for a normal approximation for $f_{Z|X}$, in which case one only needs the conditional variance function $v(x)$ for the convolution distribution in Equation (8), since $Z|X$ is zero-mean by definition. For small Z (note that the size of Z tends to decrease as inner-level sample size m increases), the general convolution expression of Equation (8) depends predominantly on $v(x)$ and less on the specific form of $f_{Z|X}$. To see this, rewrite Equation (8) as

$$
\begin{aligned}
f_{\bar{Y}}(y) &= \int_{-\infty}^{\infty} f_{Z|X}(y-x|x) f_X(x)\mathrm{d}x = \int_{-\infty}^{\infty} f_{X,Z}(x, y-x)\mathrm{d}x\\
&= \int_{-\infty}^{\infty} f_{X,Z}(y-z, z)\mathrm{d}z, && (10)
\end{aligned}
$$

and suppose we can use a second-order Taylor approximation of $f_{X,Z}(y-z, z)$ about (y, z) with respect to the first argument. Factoring $f_{X,Z}(y-z, z) = f_X(y-z) f_{Z|X}(z|y-z)$, it follows that the second-order Taylor approximation is reasonable if $f_X(y-z)$ can be approximated by a quadratic function in the vicinity of each y and if the conditional distribution $f_{Z|X}(z|y-z)$ does not vary too

strongly as its second argument varies over the vicinity of y. Notice that for each y, the vicinity of y over which the approximation must hold is determined by the effective support of the conditional distribution $f_{Z|X}(z|y)$. Because the variance of $f_{Z|X}(z|y)$ is inversely proportional to the number of inner-level replicates, the effective support of $f_{Z|X}(z|y)$ narrows as the number of inner-level replicates increases. The local quadratic approximation becomes

$$f_{X,Z}(y-z,z) \cong f_{X,Z}(y,z) - f'_{X,Z}(y,z)z + f''_{X,Z}(y,z)z^2/2, \tag{11}$$

where $f'_{X,Z}(y,z)$ and $f''_{X,Z}(y,z)$ denote the first and second derivatives of $f_{X,Z}(y,z)$ with respect to y. Substituting Equation (11) into (10) gives (see Appendix A for details)

$$\begin{aligned} f_{\bar{Y}}(y) &\cong \int_{-\infty}^{\infty} \left\{ f_{X,Z}(y,z) - f'_{X,Z}(y,z)z + f''_{X,Z}(y,z)z^2/2 \right\} \mathrm{d}z \\ &= f_X(y) + v(y)f''_X(y)/2 + v'(y)f'_X(y) + v''(y)f_X(y)/2. \end{aligned} \tag{12}$$

Equation (12) indicates that the conditional variance function $v(x)$ is the main characteristic of $f_{Z|X}$ that influences $f_{\bar{Y}}(y)$, as long as the quadratic approximation in Equation (11) is reasonable. Consequently, a normal approximation for $f_{Z|X}$ should be fairly innocuous, as long as we can obtain a reasonable estimate of $v(x)$. Towards this end, write

$$\begin{aligned} v(x) &= \mathrm{E}[Z^2|X=x] = \mathrm{E}\Big[\mathrm{E}(Z^2|\omega)|X=x\Big] = \mathrm{E}[\sigma_Z^2(\omega)|X=x] \\ &= \mathrm{E}\left[\frac{\sigma_\epsilon^2(\omega)}{m(\omega)}|X=x\right], \end{aligned} \tag{13}$$

where $\sigma_\epsilon^2(\omega) \equiv \mathrm{Var}[Y|\omega]$ denotes the conditional variance of Y with respect to all inner-level randomness, given ω for the outer-level scenario. We have added the argument ω to m to make clear that the number of inner-level replicates may depend on the outer-level scenario, e.g., if $m(\omega)$ is chosen based on the conditional mean $X(\omega)$ and/or conditional variance $\sigma_\epsilon^2(\omega)$. Different ω can result in Y having the same conditional mean $X(\omega)$ but different conditional variance $\sigma_\epsilon^2(\omega)$. The expectation in the right-most term of Equation (13) averages $\sigma_Z^2(\omega)$ across all such ω that share the same $X(\omega) = x$, weighted by the conditional distribution of $\omega|X = x$. In Section 2.3, we develop a method for estimating $v(x)$.

2.1.1 Remarks. Reference [13] used a first-order Taylor approximation of $f_{X,Z}(y-z,z)$ (with respect to its first argument) to arrive at the approximation $F_{\bar{Y}}(y) \cong F_X(y) + v'(y)F'_X(y)/2 + v(y)F''_X(y)/2$ in the context of quantifying the bias in the naïve estimator of $F_X(y)$. Integrating Equation (12) leads to the same expression. Hence, the approximation of $F_{\bar{Y}}(y)$ used by Reference [13] remains valid even when a first-order approximation of $f_{X,Z}(y-z,z)$ is invalid, as long as a second-order approximation is valid. This is important, because one might expect typical $f_X(y)$ to have non-negligible curvature and be much better approximated locally by a quadratic approximation than by a linear approximation. It should be noted that Reference [13] did not use this result to develop a bias-corrected estimator of $F_X(y)$. Rather, they used it to quantify the bias and optimally allocate n and m_i to minimize the MSE.

2.2 The QP Deconvolution Estimator for f_X

In this section, we derive the QP estimator based on the convolution Equation (8) to obtain a bias-corrected estimator of f_X from the empirical distribution of $\bar{Y}$ in the context of nested simulation. The QP density deconvolution estimator of Reference [36] produces an estimate of f_X, given a sample of n i.i.d. noisy observations $Y_i = X_i + Z_i$ $(i = 1, 2, \ldots, n)$ of X. They assumed the noise Z is additive and independent of X with a known noise pdf f_Z, and their QP estimator is based on the convolution equation $f_{\bar{Y}}(y) = (f_X * f_Z)(y) = \int_{-\infty}^{\infty} f_Z(y-x)f_X(x)\mathrm{d}x$. However, in our

nested simulation setting, the noise pdf $f_{Z|X}$ is a conditional pdf that depends on X. Furthermore, although we use a zero-mean normal distribution for $f_{Z|X}$ as discussed in the previous section, its variance function $v(x)$ is unknown and must be estimated. This section is devoted to the derivation of the QP estimator based on Equation (8), and estimation of $v(x)$ is discussed in the next section.

We estimate a discretized version of f_X over a grid of equally spaced support points $\{x_j : 1 \le j \le K\}$ for f_X and $f_{\bar{Y}}$, where $x_1 = \min\{\bar{Y}_i : 1 \le i \le n\}$, and $x_K = \max\{\bar{Y}_i : 1 \le i \le n\}$. Denote the discretized pdf values by $f_{X,j} \equiv f_X(x_j)$ for $1 \le j \le K$, and similarly for $f_{\bar{Y}}(\cdot)$. Let the K-length vectors $\mathbf{f}_X = [f_{X,1}, f_{X,2}, \ldots, f_{X,K}]^T$ and $\mathbf{f}_{\bar{Y}} = [f_{\bar{Y},1}, f_{\bar{Y},2}, \ldots, f_{\bar{Y},K}]^T$ represent the pdfs f_X and $f_{\bar{Y}}$, respectively, at the discrete locations. As an estimate of $\mathbf{f}_{\bar{Y}}$, we will use the histogram of the observations $\{\bar{y}_i : i = 1, 2, \ldots, n\}$ with bins centered at the same set of K support points. That is, the jth element of the estimator $\widehat{\mathbf{f}}_{\bar{Y}}$ of $\mathbf{f}_{\bar{Y}}$ is the histogram bin height at x_j, assuming the histogram is scaled to be in pdf units. The discretized estimator $\widehat{\mathbf{f}}_X$ of the pdf f_X is likewise represented as a K-length vector.

Defining $\delta = (x_K - x_1)/(K-1)$, the discretized version of Equation (8) can be written as:

$$\mathbf{f}_{\bar{Y}} \cong \mathbf{C}\mathbf{f}_X \iff \begin{bmatrix} f_{\bar{Y},1} \\ \vdots \\ f_{\bar{Y},K} \end{bmatrix} \cong \delta \begin{bmatrix} f_{Z|X}(x_1 - x_1|x_1) & \ldots & f_{Z|X}(x_1 - x_K|x_K) \\ \vdots & \ddots & \vdots \\ f_{Z|X}(x_K - x_1|x_1) & \ldots & f_{Z|X}(x_K - x_K|x_K) \end{bmatrix} \begin{bmatrix} f_{X,1} \\ \vdots \\ f_{X,K} \end{bmatrix}, \tag{14}$$

where the elements of the convolution matrix $\mathbf{C}$ are determined from the conditional noise distribution $f_{Z|X}$, as discussed in Section 2.1. Estimation of $f_{Z|X}$ is described in Section 2.3.

The basic QP density deconvolution formulation is:

$$\begin{aligned} \widehat{\mathbf{f}}_X \quad &= \quad \underset{\mathbf{f}_X}{\mathbf{argmin}} \left[\|\widehat{\mathbf{f}}_{\bar{Y}} - \mathbf{C}\mathbf{f}_X\|^2 + \lambda Q(\mathbf{f}_X) \right] \\ s.t. \quad & \quad \delta \mathbf{1}^T \mathbf{f}_X = 1 \\ & \quad \mathbf{f}_X \ge \mathbf{0}, \end{aligned} \tag{15}$$

where $Q(\mathbf{f}_X)$ is a regularization term, and λ is a regularization parameter. We use second-derivative regularization $Q(\mathbf{f}_X) = \|\mathbf{D}_2\mathbf{f}_X\|^2$, where $\mathbf{D}_2$ is an appropriately defined second-order difference matrix operator. That is, we penalize large second derivatives of f_X. Some form of regularization is essential in these types of deconvolution problems, because the convolution matrix $\mathbf{C}$ is typically poorly conditioned, in which case an unregularized solution to Equation (15) would typically be oscillatory or otherwise badly behaved. Note that the unregularized ($\lambda = 0$) solution to Equation (15) requires inversion of the $\mathbf{C}$ matrix. See Reference [36] and the references therein for additional discussion of this issue. The vector $\mathbf{1}$ is a column vector of ones, and $\mathbf{f}_X \ge \mathbf{0}$ means that all elements of $\mathbf{f}_X$ are nonnegative. The regularization parameter λ can be selected via a SURE-like method, as discussed in Reference [36]. **(SURE is Stein's Unbiased Risk Estimator.)** The SURE-like method in Reference [36] is derived in the context of homoscedastic noise that is independent of X. However, it remains valid in the nested simulation context, i.e., the noise Z is conditionally (given X) heteroscedastic. This is because the only assumption for the derivation is that $\widehat{\mathbf{f}}_{\bar{Y}}$ follows a multinomial distribution with mean vector equal to $\delta\mathbf{C}\mathbf{f}_X$, which still holds with conditionally (given X) heteroscedastic noise as $\widehat{\mathbf{f}}_{\bar{Y}}$ is obtained from the histogram. In addition to the automated SURE-like method, Reference [36] also discussed a graphical method for selecting λ. The former is convenient if the analysis must be repeated many times (e.g., when analyzing the performance of the approach in a Monte Carlo setting), whereas the latter has advantages from a more practical user's perspective in which the analysis is conducted a single time.

In addition, a number of common features of density functions can be translated into shape constraints on $\mathbf{f}_X$, and incorporated into Reference (15). Specifically, many shape constraints can

be formulated as a linear constraint, $\mathbf{A}_t\mathbf{f}_X \geq \mathbf{0}$, on f_X for some matrix $\mathbf{A}_t$, where the subscript t is the index of the shape constraint. Such shape constraints include tail monotonicity, tail convexity, and unimodality. Moreover, if there is information on the support of f_X, e.g., that we know the support of f_X lies within the interval $[x_a, x_b]$ for some specified $x_1 \leq x_a < x_b \leq x_K$, then we could reformulate Reference (15) by replacing the K-dimensional $\mathbf{f}_X$ by its reduced $(b-a+1)$-dimensional counterpart $[f_{X,a}, f_{X,(a+1)}, \ldots, f_{X,b}]^T$ and also replacing the $K \times K$ matrix $\mathbf{C}$ by its $K \times (b-a+1)$ counterpart comprised of columns $\{a, a+1, \ldots, b\}$ of $\mathbf{C}$. It has been demonstrated in Reference [36] that including any such "prior" knowledge we may have regarding $\mathbf{f}_X$ can improve the pdf estimator substantially.

2.3 Estimating the Conditional Error Variance Function

As discussed in Section 2.1, the conditional error distribution $f_{Z|X}(z|x)$ required in the deconvolution estimator is taken to be a zero-mean normal distribution, so only its conditional error variance function $v(x)$ in Equation (13) must be estimated. The QP method of Reference [36] assumes known homoscedastic (independent of X) error distribution, whereas in two-level simulation settings the error distribution is unknown, and it is often strongly heteroscedastic with $v(x)$ depending strongly on x. In this section, we provide an approach to estimate $v(x)$.

To make the problem more tractable, suppose that $m(\omega) = m(X(\omega))$ depends on ω only via $X(\omega)$, which obviously holds if $m(\omega) = m$ is constant. Then

$$v(x) = \mathrm{E}\left[\frac{\sigma_\epsilon^2(\omega)}{m(X(\omega))}|X(\omega) = x\right] = \frac{h(x)}{m(x)},$$

where $h(x) \equiv \mathrm{E}[\sigma_\epsilon^2(\omega)|X(\omega) = x]$. Conditioned on ω_i, the sample variance

$$S_i^2 \equiv \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(Y_{ij} - \bar{Y}_i)^2$$

is an unbiased estimator of $\sigma_\epsilon^2(\omega_i)$. Thus,

$$\mathrm{E}[S_i^2|X(\omega_i) = x] = \mathrm{E}\Big[\mathrm{E}[S_i^2|\omega_i]|X(\omega_i) = x\Big] = \mathrm{E}[\sigma_\epsilon^2(\omega_i)|X(\omega_i) = x] = h(x). \tag{16}$$

In light of Reference [16], hypothetically supposing X_i were known (e.g., $X_i \cong \bar{Y}_i$ for large m_i), we could use a loess smoother of the scatterplot of S_i^2 vs. X_i $(i = 1, \ldots, n)$ (or perhaps the square of a loess smoother of the sample standard deviation S_i vs. X_i) as an estimator of the function $h(x)$. Note that "loess" stands for "locally weighted estimation and scatter plot smoothing" and is a general nonparametric regression approach that is commonly used to smooth low-dimensional scatter plots [6, 14]. (A smoothing spline could be used instead of loess and would give a similar estimate.) In practice, however, if we use $\bar{Y}_i$ as a surrogate for X_i and fit a loess smoother to the scatterplot of S_i^2 vs. $\bar{Y}_i$, then what we estimate is the function $g(x) \equiv \mathrm{E}[\sigma_\epsilon^2(\omega)|\bar{Y} = x]$ rather than the desired function $h(x) = \mathrm{E}[\sigma_\epsilon^2(\omega)|X = x]$. For large m_i, for which $X_i \cong \bar{Y}_i$, this does not pose a serious problem. However, when m_i is small, and as argued before a small m_i is desirable, the functions g and h can be quite different, and using an estimate of g in place of the desired estimate of h can adversely affect the QP estimator. We demonstrate this later in the examples.

The problem of g differing from h is similar to but much more complicated than the well-known "errors-in-predictors" problem in linear regression. In simple linear regression with homoscedastic errors in the predictor and in the response, there is a simple correction to the slope of the fitted line to give an unbiased estimate of the desired regression coefficient. In our case, the observed predictor variable for the loess fit is $\bar{Y} = X + Z$ instead of X, and the error in $\bar{Y}$ is heteroscedastic, depending on the randomness inherent to both the outer-level and inner-level replicates. Moreover,

our estimate of h is nonparametric. Although we omit the results for brevity, we have investigated various approaches for correcting for the errors-in-predictors phenomenon for heteroscedastic nonparametric regression, e.g., using deconvolution-like approaches for recovering $h(x)$ from $g(\bar{y})$. However, none of the approaches performed robustly.

Instead, to obtain a reliable estimate of the h function, we adopt the approach of *removing* the errors-in-predictors phenomenon, as opposed to correcting for it. Specifically, we run additional inner-level replicates for a relatively small subset of outer-level replicates, so $\bar{Y}$ will be close to X on these replicates. Then, we estimate $h(x)$ directly by fitting a loess model to only these fewer, but higher-quality, data points. Let $\{Y^a_{i,j} : i = 1, 2, \ldots, n^a; j = 1, 2, \ldots, m^a_i\}$ denote the additional data, where n^a and m^a_i denote the numbers of additional outer-level and inner-level replicates, respectively. For the ith additional outer-level replicate, let $\bar{Y}^a_i$ and S^a_i denote its sample average and sample standard deviation (across the m^a_i inner-level replicates), respectively. The loess model will be fit with $\{\bar{Y}^a_i : i = 1, \ldots, n_a\}$ as the predictor variable and $\{S^a_i : i = 1, \ldots, n_a\}$ as the response variable (which gives an estimate of the square root of $h(x)$). The additional data are generated using the following procedure, which is implemented after having run the main nested simulation that generated the original data $\{Y_{i,j} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, m_i\}$ and their corresponding averages $\{\bar{Y}_i : i = 1, \ldots, n\}$:

(1) We begin with $n_a = 12$ (which is increased as described later, if needed), and the initial 12 additional outer-level replicates are taken to be those corresponding as closely as possible to 12 evenly spaced points over the interval $[\bar{Y}_{min}, \bar{Y}_{max}]$, i.e., to $\{\bar{Y}_{min} + \frac{i-1}{11}(\bar{Y}_{max} - \bar{Y}_{min}) : i = 1, 2, \ldots, 12\}$, where $\bar{Y}_{min}$ and $\bar{Y}_{max}$ denote the minimum and maximum of $\{\bar{Y}_i : i = 1, \ldots, n\}$. This gives an initial set of additional outer-level replicates that roughly span the sample space of $\{X_i : i = 1, \ldots, n\}$, to ensure that the predictor values for the loess fit are spread across the entire range.

(2) For each of the n_a outer-level replicates, initially conduct $m^a_i = 100$ inner-level replicates, and compute the averages $\{\bar{Y}^a_i : i = 1, \ldots, n_a\}$ and standard deviations $\{S^a_i : i = 1, \ldots, n_a\}$.

(3) For each of the n_a outer-level replicates, sequentially add inner-level replicates (i.e., increase m^a_i) until the standard error $S^a_i/\sqrt{m^a_i}$ of $\bar{Y}^a_i$ is below 3% of the range $\bar{Y}_{max} - \bar{Y}_{min}$. The purpose of this step is to ensure that each additional $\bar{Y}^a_i$ is close enough to X_i that the errors-in-predictors can be ignored.

(4) Determine whether the number n_a of additional outer-level replicates is sufficient to obtain a reliable loess estimate for h, and if not, select another additional outer-level replicate at which to run additional simulations:

(i) Fit a loess model with $\{\bar{Y}^a_i : i = 1, \ldots, n_a\}$ as the predictor variable and $\{S^a_i : i = 1, \ldots, n_a\}$ as the response variable. For any x in the domain of the loess model, denote the fitted response value and its standard error by $l(x)$ and $s_l(x)$, respectively. If using the **loess** function in R ([27]), which we have used in all of our examples, then $s_l(x)$ is produced via the `predict.loess` command.

(ii) Compute $l_{max} \equiv \max_{\bar{Y}^a_{min} \le x \le \bar{Y}^a_{max}} l(x)$, $x^* \equiv \text{argmax}_{\bar{Y}^a_{min} \le x \le \bar{Y}^a_{max}} s_l(x)$, and $s^* \equiv s_l(x^*)$, where $\bar{Y}^a_{min}$ and $\bar{Y}^a_{max}$ denote the minimum and maximum, respectively, of $\{\bar{Y}^a_i : i = 1, \ldots, n_a\}$. If $s^* \ge 0.03 \times l_{max}$ (i.e., if the maximum standard error exceeds 3% of the largest predicted response value), then add another outer-level replicate sampled independently and increase n_a by one. As the newly added outer-level replicate, take the replicate corresponding to whichever of the remaining $\{\bar{Y}_i : i = 1, \ldots, n\}$ is closest to x^* but within the interval $[\bar{Y}^a_{min}, \bar{Y}^a_{max}]$. Otherwise, if $s^* < 0.03 \times l_{max}$, then stop.

(iii) Repeat Steps (2) and (3), but only for the newly added outer-level replicate.

(5) Repeat Step (4) until the criterion $s^* < 0.03 \times l_{max}$ is satisfied.

The above procedure is intended to produce set of high-quality data points for estimating the function $h(x)$, in the sense that the sample standard deviations S_k and sample means $\bar{Y}_k$ should be close to the true standard deviations $\sigma_\epsilon(\omega_k)$ and the true means X_k, respectively. At termination, the square of the loess fit $l(x)$ in Step (4i) is taken as the estimate of the $h(x)$ function. We denote the estimate as g_a, where the subscript a stands for additional simulations. We have chosen to fit the loess smoother to S_k and use the square of the smoother, as opposed to fitting directly to S_k^2, because our local linear smoother implementation extrapolates linearly, and $\sqrt{h(x)}$ is more likely to be approximately linear in x over the boundary regions than is $h(x)$ (i.e., the standard deviation is more likely to be proportional to the mean than is the variance). Note that this only applies to extrapolation outside the range of data, if needed. Our estimator of $h(x)$ within the data range is nonparametric. The performance of g_a and its effect on the corresponding QP pdf estimator will be illustrated in Section 3.

The 3% threshold values in Steps (3) and (4-ii) could be adjusted as tuning parameters to increase or decrease the number of additional inner-level and outer-level replicates, if desired. We have found the above 3% values to be reasonable and to work robustly for all of the examples we have tried. To help decide whether a sufficient number of additional inner-level and outer-level replicates have been conducted, we recommend plotting $g_a(\bar{y})$ on top of a scatter plot of S_i^2 vs. $\bar{y}_i$ (for both the original replicates and the additional replicates), as in Figure 2(a). If the loess-fitted $g_a(\bar{y})$ appears badly behaved (e.g., "wiggly"), this is an indication that additional replicates are needed.

2.4 Consistency and Rate of Convergence

Appendix D addresses some asymptotic properties of the QP deconvolution estimator. Appendix D.1 contains a formal proof that the QP deconvolution estimator of F_X is consistent. Appendix D.2 contains a heuristic argument that the convergence rate of $\hat{f}_X$ is the same as the rate of the jackknife estimator in Steckley et al. [28].

In Appendix D.1 it is first shown that the solution to Equation (15), if transformed as in Equation (8), is strongly $\mathcal{L}_2$-consistent for $f_{\bar{Y}}$. Alone, this is not particularly interesting (the histogram estimate itself is strongly $\mathcal{L}_2$-consistent for $f_{\bar{Y}}$), but the form of the estimator in Equation (15) allows us to strengthen this result to a weak consistency result for F_X in Theorem 3. It says, in abbreviated form here:

Theorem 1. *Let $\widehat{F}_{X,n}$ and F_X be the CDFs of $\widehat{f}_{X,n}$ and f_X, respectively. Under certain conditions, with probability 1, $\widehat{F}_{X,n}(t) \to F_X(t)$ for all t.*

In fact, since F_X is absolutely continuous, it follows that this convergence is uniform (see Lemma 2.11 of Reference [33]). That is, with probability 1, $\sup_t |\widehat{F}_{X,n}(t) - F_X(t)| \to 0$. Furthermore, by Lemma 21.2 of Reference [33], the quantile functions converge weakly as well. Letting $F^{-1}(p) = \inf\{t : F(t) \geq p\}$ be the quantile function of F, Theorem 3 implies that with probability 1, $\widehat{F}_{X,n}^{-1}(p) \to F_X^{-1}(p)$ for all p.

3 NUMERICAL EXAMPLES AND DISCUSSION

3.1 Examples with Uniform Allocation

In this section, we illustrate the performance of our bias-corrected QP estimator and compare it with the Steckley et al. estimator [28] and the naïve pdf estimator on two nested simulation examples. One is a simple parametric version of the portfolio loss example discussed in the introduction, which we use to demonstrate certain properties of our proposed QP estimator. The second example is a more complex call center simulation example. In both examples, our focus is on the output

analysis of the simulation results rather than optimizing the choice of the number of the outer and inner-level replicates. Here, we assume a uniform allocation, i.e., $m_i = m$ for all $i = 1, 2, \ldots, n$.

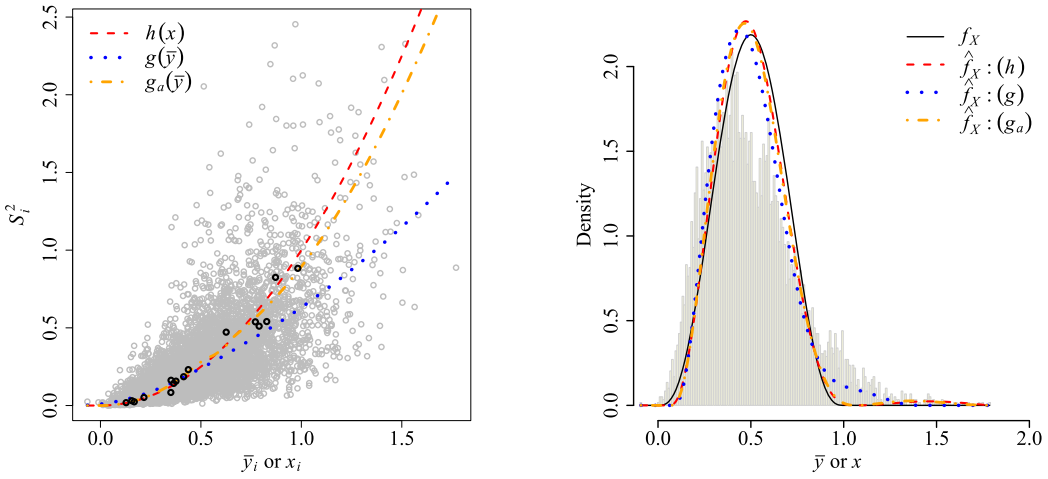
3.1.1 Portfolio Loss Example. In a variant of the portfolio loss example discussed in the introduction (also see Reference [13] for a more detailed description of a typical example and two-level simulation framework), suppose we want to estimate the fractional portfolio loss (loss relative to the portfolio value) given default, using two-level simulation. In the following, we will refer to the fractional portfolio loss given default as "portfolio loss" for simplicity. In real life, at the horizon time, the portfolio will be priced by conducting a single-level simulation that corresponds to the inner level of our nested simulation. This simulation will simulate what will happen between the horizon time and the final maturity time, beginning with the single state in which real life finds itself at the horizon time. We need an outer-level simulation, because the purpose of our nested simulation is to assess at the current time the risks of realizing large portfolio losses in the future, at the horizon time. This is done in our simulation by estimating the distribution of portfolio prices at the horizon time. This involves simulating what might happen between the current time and the horizon time, which is what the outer level of the nested simulation does.

Using the same notation in the introduction and Equations (3)–(7), the portfolio loss for the ith scenario ($i = 1, \ldots, n$) generated on the ith replicate of the outer level is $X_i = X(\omega_i)$. Given the outer-level random outcome ω_i for that scenario, the inner level of the simulation then generates the loss-related quantities $\{Y_{ij} : j = 1, 2, \ldots, m\}$ across the m inner-level replicates, such that their average $\bar{Y}_i = m^{-1} \sum_{j=1}^{m} Y_{ij}$ serves as a noisy estimate of the loss X_i.

In practice, the risk factors embedded in ω_i are nonparametric or extremely high dimensional (e.g., the entire price and cash flow histories of all positions in the portfolio). However, to have a known analytical distribution for X, to which we can compare the various estimators in this example, we conduct the nested simulation as follows: We simulate in the ith outer-level replicate a portfolio loss $X_i \sim \text{Beta}(4, 4)$ whose mean equals 0.5. Fractional portfolio loss given default is commonly modeled as a beta distribution References ([1, 11, 17]). In the inner-level replicate, we simulate the mean-zero pricing error $\epsilon_{ij}|X_i = x_i \sim \text{Gamma}(4, 2/x_i) - 2x_i$ for $j = 1, \ldots, m$. The simulated average portfolio losses are $\bar{Y}_i = X_i + Z_i$, for $i = 1, \ldots, n$, where $Z_i = m^{-1} \sum_{j=1}^{m} \epsilon_{ij}$ depends on X_i. Under this parameterization, the conditional variance is $\text{Var}[Z|X = x] = x^2/m$, and the signal-to-noise ratio can be analytically obtained as $\text{Var}[X]/\text{Var}[Z] = m\text{Var}[X][\text{Var}[X] + \text{E}^2[X]]^{-1} = m/10$ (see Appendix B).

As discussed in Section 2.3, when m is small, the difference between h and g has an adverse effect on the QP estimator. Using g_a as an estimator of h, rather than the original loess estimator g, improves the estimation of h and the resulting QP pdf estimator. For example, Figure 2(a) shows the scatterplot of S_i^2 vs. $\bar{Y}_i$ ($i = 1, 2, \ldots, n$) for a typical replicate of the above nested simulation using a total computational budget $M = nm = 5 \times 10^4$ and the allocation $m = 8$ and $n = 6,250$. It also shows the true h function, which is $h(x) = x^2$ under the parameterization of the portfolio loss example, along with the loess estimates g and g_a. We can see that g_a, which only consumes about 3% additional computational budget, substantially improves upon the original loess estimate g. Figure 2(b) shows the corresponding QP estimators using the functions h, g and g_a from Figure 2(a). The QP estimator using g_a is better than the one using g, both in the middle quantiles and in the upper tail. The QP estimator using h is only shown as a reference benchmark, since h is unknown in practice. Notice that the QP estimator using g_a is almost as good as the one using h. In the subsequent discussions, unless noted otherwise, the QP estimator will mean the one using g_a.

Figure 1 compares the performances of the bias-corrected QP deconvolution estimator, the bias-corrected estimator developed in Reference [28], and the naïve estimator (a kernel density estimator) for a typical replicate of the nested simulation under the allocation $m = 5$, $n = 10^4$ and



(a) Estimation of h function

(b) QP pdf estimators

Fig. 2. Performance of the proposed $h(x)$ function estimation procedure and the corresponding QP estimators for the portfolio loss example for a total computational budget of $M = nm = 5 \times 10^4$ and the allocation of $m = 8, n = 6,250$. Panel (a) is the scatterplot of the sample variance against the sample mean, along with the true function h (red dashed) and the estimates g (blue dotted) and g_a (orange dot-dashed). Panel (b) shows the corresponding QP estimators using h, g, and g_a, respectively. The curves for h and g_a are difficult to distinguish because they are so close.

demonstrates better performance for the QP deconvolution estimator. For fair comparisons, we slightly increased the budgets for the naïve and the Steckley et al. estimators so they had the same budget as the QP estimator with the additional replicates for estimating g_a, which typically increased the budget by about 3% (the exact percentage varies from replicate to replicate). We can see from Figure 1 that the histogram and the naïve estimator are overly dispersed relative to the true f_X, which suggests that a small number of inner-level replicates ($m = 5$ in this case) introduces substantial bias when we use the naïve estimator. The Steckley et al. estimator, although better than the naïve estimator, does not correct the bias enough for this small m. In contrast, the QP estimator with only the two universally applicable shape constraints (i.e., integrate-to-one and nonnegativity) is much closer to the true density across the entire domain of the distribution, including the tails. The L_1 distances between the true pdf and the pdf estimators (defined as $L_1(\widehat{f_X}, f_X) = \int |\widehat{f_X}(x) - f_X(x)|\mathrm{d}x$) for the QP estimator, the Steckley et al. estimator and the naïve estimator are 0.152, 0.364, and 0.523, respectively, for this typical replicate shown in Figure 1. In fact, we have selected this replicate as typical, because each of the three estimators had an L_1 distance that is fairly close to their median L_1 distance across 30 replicates. The median L_1 distances for the QP, Steckley et al. and naïve estimators are 0.152, 0.368, 0.524, respectively.

Moreover, Figure 3 illustrates that when the total computational budget is fixed, the QP estimator outperforms the naïve and the Steckley et al. estimators at their respective optimal allocations. The optimal allocations were determined by finding the pair (n, m) that results in the minimum aggregate MSE measure, as described in the next paragraph. Figure 3 compares the three estimators for a typical replicate at their optimal allocations under a total computation budget $M = 5 \times 10^4$. The optimal allocations are $m = 22$ for the QP estimator, $m = 175$ for the naïve estimator, and $m = 100$ for the Steckley et al. estimator. Since the total computational budget is fixed, a larger number of inner-level replicates means a fewer number of outer-level replicates, i.e., fewer

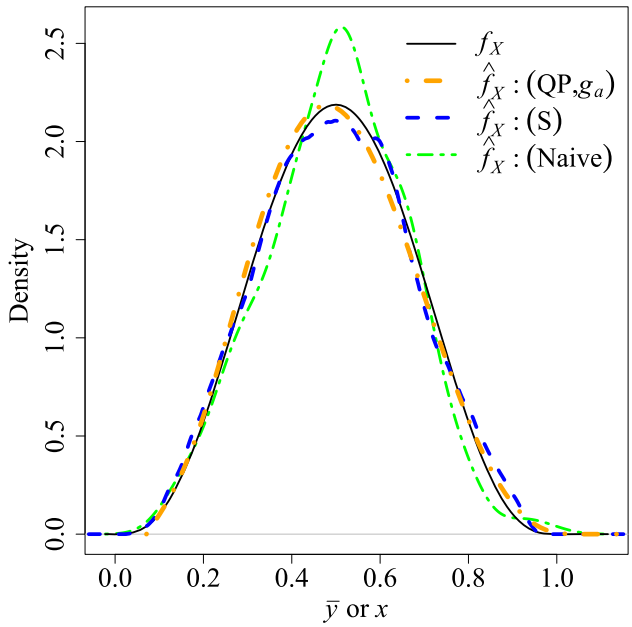


Fig. 3. pdf estimation results for the same portfolio loss example in Figure 1 but with optimal allocations for each method, which are $m = 22$ for the QP estimator, $m = 175$ for the naïve estimator, and $m = 100$ for the Steckley et al. estimator. All methods are under the same total computational budget $M = 5 \times 10^4$. Comparison with Figure 1 shows that the QP estimator performs more robustly to choice of m, which is important, since the optimal m is difficult to determine in practice.

observations of $\bar{Y}$, which causes the naïve estimator to perform less reliably. Comparing the two bias-corrected estimators, the QP estimator is better than the Steckley et al. estimator for this typical replicate in Figure 3. The L_1 distances of the QP estimator, the Steckley et al. estimator, and the naïve estimator are 0.0489, 0.0550, and 0.111, respectively, for this typical replicate shown in Figure 3. Notice that the performance of our QP estimator for smaller m (Figure 1), relative to its performance for optimal m (Figure 3), does not degrade nearly as much as does the performance of the naïve estimator and the Steckley et al. estimator. We investigate this further below.

To more quantitatively compare the methods, we conducted a **Monte Carlo (MC)** simulation with fixed computational budget at various allocations. We vary m to control the signal-to-noise ratio ($\text{Var}[X]/\text{Var}[Z]$) to vary from 0.8 to over 20. To distinguish the performance in estimating the tails vs. the middle quantiles of the pdfs, we consider the MSE for estimating nine quantiles corresponding to probabilities $p \in \{0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99\}$ for each allocation. The MSEs of the quantile estimates are displayed in Tables 1(a) and 1(b), which correspond to the total budgets $M = 5 \times 10^4$ and $M = 10^6$, respectively. On each MC replicate, we generated a new set of M $Y_{i,j}$ observations for each allocation and then applied the corresponding pdf estimators to the data. The regularization parameter for each MC replicate was selected automatically via the SURE-like method [36]. The number of MC replicates was chosen to ensure that the standard errors of the MSEs are about 1% the MSE values. The MSEs reported in the tables for each estimator are for their optimal allocations for the given computational budget. The optimal allocations were chosen to minimize the aggregate MSE measure defined as $\sum_{i=1}^{9} MSE_i/[p_i(1-p_i)]$, where $MSE_i = MSE(\hat{p}_i)$ is the MSE for the ith probability estimator, among 10 different allocations. The optimal numbers of inner-level replicates for the given total computational budget for each estimator are displayed

Table 1. Comparisons of Quantile Estimation MSEs ($\times 10^5$) for the Portfolio Loss Example for Various Probabilities (p) under the Optimal Allocations

(a) Fixed total budget $M = 5 \times 10^4$

p	QP (h) ($m = 12$)	QP (g_a) ($m = 22$)	Steckley ($m = 100$)	Naïve ($m = 175$)
0.01	0.638	0.579	1.68	6.22
0.05	3.67	3.79	8.90	24.0
0.10	4.81	8.00	17.3	40.7
0.25	8.93	17.4	38.3	69.2
0.50	22.90	24.4	53.7	71.2
0.75	12.2	21.7	44.0	55.2
0.90	8.41	16.5	28.6	49.4
0.95	10.1	12.9	21.3	42.0
0.99	6.86	7.01	9.57	20.6

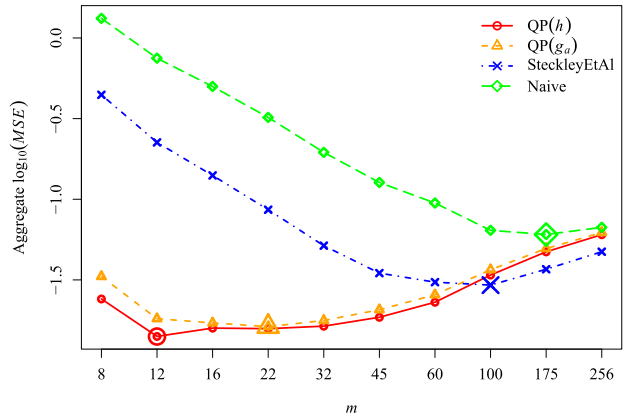
(b) Fixed total budget $M = 10^6$

p	QP (h) ($m = 60$)	QP (g_a) ($m = 60$)	Steckley ($m = 175$)	Naïve ($m = 256$)
0.01	0.124	0.146	0.139	0.888
0.05	0.397	0.372	0.802	3.430
0.10	0.631	0.666	1.360	5.730
0.25	1.180	1.380	3.250	9.330
0.50	1.730	1.850	4.790	9.140
0.75	1.470	1.840	3.860	6.030
0.90	1.080	1.170	2.320	6.300
0.95	0.993	1.070	1.810	6.370
0.99	0.549	0.635	0.781	3.410

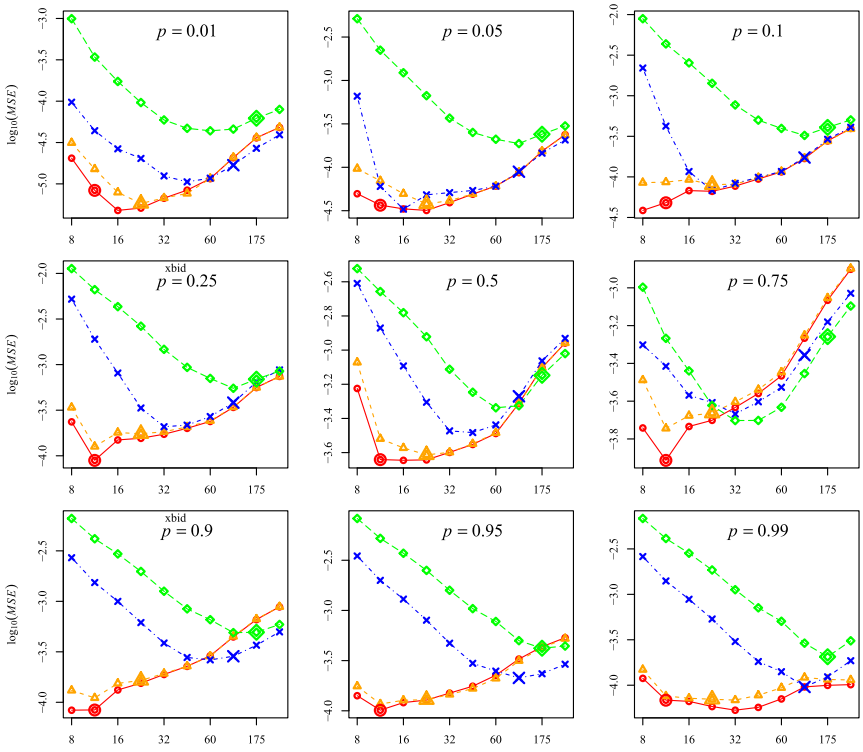
Note: (optimal m values shown in parentheses) for Each of the Four pdf Estimators. The Regularization Parameters Were Selected Using Automated Methods.

in parentheses in the first rows of the tables. We see from Tables 1(a) and 1(b) that the *MSEs* for the QP estimator are about half of those for the bias-corrected Steckley et al. estimator across all quantiles and between about two to ten times smaller than for the naïve estimator (which we take to be the histogram).

Figure 4 provides a more complete perspective by plotting the aggregate MSE measure and the MSE measures at nine separate quantiles for each estimator at ten allocations (i.e., 10 values of m and $n = M/m$) under fixed computational budget $M = 5 \times 10^4$. The aggregate MSE (Figure 4(a)) conveys the overall trends of how the performances of the different pdf estimators change when the allocation m and $n = M/m$ varies. Figure 4(b) shows the MSE measure at nine quantiles separately. The highlighted bullets indicate the optimal allocation for each estimator based on the aggregate MSE measure. In practice, if all quantiles are of interest, then the same allocation must be used for each quantile. Consequently, the MSE at the highlighted bullets is of particular interest for comparison purposes. The better robustness to choice of m that is seen by comparing Figures 1 and 3 is substantiated by Figure 4, which shows that the performance of our QP estimator does not degrade as much as the performance of the naïve estimator and the Steckley et al. estimator



(a) Comparison of the aggregate MSEs under various allocations.



(b) Comparison of MSEs of nine separate quantiles.

Fig. 4. Comparisons of the performance of four pdf estimators for the portfolio loss example under fixed budget $M = 5 \times 10^4$ and at various inner level allocations ($m = 8, 16, 24, 32, 64, 128, 256$). The x-axis in all figures represents the inner level allocation m. The four estimators are the naïve estimator (green long-dashed), the Steckley et al. estimator (blue dot-dashed line), the QP estimator with g_a (orange dashed), and h (red solid), respectively. The vertical axis is the MSE measure in the log-scale of base 10. The highlighted bullets (i.e., the single enlarged symbol for each curve) indicate the optimal allocation for each estimator based on the aggregate MSE measure shown in Figure 4a.

as m differs from its optimal value, especially for the tail quantiles. This is likely because it does a better job at correcting for the over-dispersion bias than the Steckley et al. estimator, and the naïve estimator does not even attempt to correct for this bias. Robustness to m is important, because it is difficult to determine the optimal m in practice. Moreover, there is no single optimal m if multiple features of f_X are of interest, since the optimal value of m for estimating quantiles depends heavily on which quantile, as evident from Figure 4.

As shown in Figure 4, the QP estimator with g_a performs comparably to the QP estimator with h even when $m \leq 45$, which suggests that the estimator g_a using the procedure in Section 2.3 is a reasonable estimator of the true h function. Moreover, the QP estimator outperforms the Steckley et al. estimator and the naïve estimator for most of the 10 different allocations for this example, especially when m is small. Although the Steckley et al. method is slightly better than the QP method for large inner-level sample sizes ($m \geq 60$), and both the QP and Steckley et al. estimators are close to the naïve histogram estimator when m grows, the QP estimator achieves much better MSE results for small to moderate m ($m = 8$ to 45).

As mentioned above, if we compare the performances at only the highlighted bullets, i.e., at each estimator's optimal allocation, then the QP estimator performs much better than the alternatives for almost all nine quantiles and also for the aggregated MSE. Another interesting phenomenon is that when the same large m value (e.g., $m \geq 60$) is used for both the Steckley et al. and the QP estimator, the former performs a little better overall, more noticeably so for the upper quantiles. This is understandable, considering that the optimal m values are substantially larger for the Steckley et al. method than for the QP method. A potential explanation for this is that the QP estimator is more discretized, whereas the Steckley et al. estimator is a smoother kernel-based method. Thus, the latter may have advantages when data are sparser in the tail regions, which happens when m is large (since the computational budget is fixed, large m means smaller n and fewer $\bar{Y}$ observations). For example, for large m the QP estimator may become truncated at some tail quantiles due to lack of data, whereas the Steckley et al. estimator tends to be smoother and more continuous. However, it should be emphasized that at their optimal allocations (either using the single m optimized for the aggregate MSE measure or using separate m optimized for each quantile) the QP estimator still substantially outperforms the Steckley et al. estimator and the naïve estimator.

Regarding the assumption of normality for Z, it should be noted that the errors themselves do not need to be normally distributed, as long as the average Z of m errors is close to normal. In the portfolio loss example, the errors follow a gamma distribution (but translated, to have zero mean) with shape parameter 4, which is a right-skewed distribution. However, even for relatively small m, the distribution of Z is close to normal. To investigate this, consider a more extreme version of non-normality in which the errors follow a zero-mean version of a chi-square distribution with two degrees of freedom, denoted $\epsilon_{i,j} \sim \chi^2_2$. Note that the χ^2_2 distribution is also an exponential distribution, which is highly right-skewed with a jump discontinuity at zero. Figure 5 plots a standardized version (translated to have zero mean and scaled to have unit variance) of the χ^2_2 pdf as the curve for $m = 1$, since $Z = \epsilon$ for $m = 1$. The figure also plots the standardized pdfs for Z for various $m \in \{1, 2, 5, 10, 20, 50, 200, \infty\}$, which get progressively closer to the standard normal distribution as m increases. Note that by the reproductive property of the chi-square distribution, the distribution of Z is χ^2_{2m}. Even though the distribution for $\epsilon_{i,j}$ is highly non-normal, the distribution for Z is close to normal for around $m = 10$ or larger. We anticipate that for most stochastic two-level simulations, the errors will not differ from normality any more than an exponential distribution differs from normality. For example, we do not envision many applications in which the $Y_{i,j}$ are Bernoulli. If one encounters an application in which the errors are that far from normal, then one should use caution to ensure that m is large enough to give an approximately normal distribution for Z.

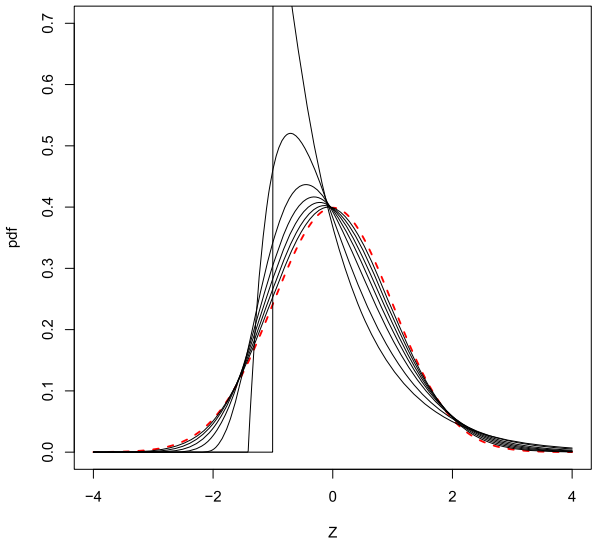


Fig. 5. Illustration of the effect of m on the approximate normality of Z for a highly non-normal error distribution. The red dashed curve is the standard normal pdf, and the other curves are pdfs for (a standardized version of) Z for $m \in \{1, 2, 5, 10, 20, 50, 200, \infty\}$, which get progressively closer to the standard normal distribution as m increases. The distribution for $\epsilon_{i,j} \sim \chi^2_2$ is the distribution for Z for $m = 1$, and the normal distribution is for $m = \infty$.

We note that the deconvolution approach is still valid and can still remove the over-dispersion bias when Z is non-normal, as long as the correct pdf $f_{Z|X}$ is used in the convolution/deconvolution Equation (8). In fact, it is well-known in the deconvolution literature that super-smooth noise distributions like Gaussian are notoriously more difficult to deconvolve than other less smooth noise distributions; see Reference [10]. It is primarily to avoid the difficulty of estimating the noise distribution that we have taken advantage of the fact that Z is an average of m errors to approximate $f_{Z|X}$ as normal in the deconvolution equations. Estimation of a non-normal $f_{Z|X}$ would be substantially complicated by the fact that it may depend on X.

3.1.2 Input Uncertainty Analysis and a Call Center Simulation Example. As a special case of our general model in which ω is parametric, consider the problem of studying the effects of uncertainty in the parameters of the distributions of stochastic inputs in a single-level stochastic simulation. For example, suppose we have an $M/M/c$ queuing system simulation in which customer arrivals follow a Poisson process whose mean arrival rate parameter $\Lambda = \Lambda(\omega)$ is fixed (i.e., constant) but unknown, and the uncertainty in Λ (due to lack of knowledge of Λ; not because it varies over time) is represented by treating it as a random variable (r.v.) that follows some specified distribution. Although the queuing system simulation of interest is a single-level Monte Carlo simulation, one typically treats the input uncertainty problem via a two-level simulation in which the ith outer-level replicate consists of generating one random draw of Λ from its distribution ($\Lambda_i = \Lambda(\omega_i)$). Each inner-level replicate comprises one run of the actual simulation model for that particular Λ_i. That is, on the jth inner-level replicate of the ith scenario, the number of customers being served during a given time period is simulated and denoted by $Y_{ij} = Y(\Lambda_i, \xi_{i,j})$, where $\xi_{i,j}$ determines all sources of randomness in the inner-level replicate due to the random Poisson arrivals (but with Poisson arrival rate parameter Λ_i fixed) and the random service times. A common performance metric is the expected number of customers served during the given time period, i.e., the conditional

expectation $X = X(\Lambda) = \mathrm{E}[Y|\Lambda]$, which is a function of the unknown Λ. Hence, the pdf f_X of the conditional expectation $X(\Lambda)$ with respect to the distribution of Λ represents the uncertainty in the system performance due to uncertainty in the input parameter Λ. As a slightly different version of this setting, suppose one computes an updated forecast of Λ the night before some target period. Assuming the forecast is not perfect and involves some uncertainty, the distribution for Λ that one would use could be some distribution with mean equal to the forecast and standard deviation equal to the standard error of the forecast. In this case, our computed distribution for the conditional expectation $E[Y|\Lambda]$ would represent the distribution of the expected number of customers served on the target day, where the distribution is with respect to the uncertainty in the forecasted Λ.

For this parametric input uncertainty problem with low-dimensional Λ, Reference [35] viewed the parameters $\Lambda = \Lambda_{true}$ as fixed but unknown and developed a Bayesian credible interval for $\mathrm{E}[Y|\Lambda_{true}]$ with uncertainty in Λ_{true} estimated via standard Bayesian methods applied to a sample of real input data. Their approach is based on using a surrogate model to represent $\mathrm{E}[Y|\Lambda]$ as a function of Λ, which tends to smooth out the noise that results from using a finite number of inner-level replicates and in this manner partially accounts for the over-dispersion problem. Although their objective of obtaining a credible interval for $\mathrm{E}[Y|\Lambda_{true}]$ differs from our objective of estimating the distribution of $\mathrm{E}[Y|\Lambda]$, one might consider adapting a similar Bayesian approach for our objective. Instead, here, we apply our deconvolution-based approach, partly because it is more general and applies to parametric or nonparametric uncertainty (the surrogate modeling in Reference [35] requires a relatively low-dimensional parameterization), and also because a second Bayesian mechanism in Reference [35] would exacerbate the over-dispersion problem if no surrogate modeling is used. In follow-up work, Reference [3] notes that surrogate modeling approaches are not effective for high-dimensional parametric uncertainty in input distributions or for nonparametric empirical input distributions and developed a bootstrapping **confidence interval (CI)** that takes into account nonparametric input uncertainty but is not unduly widened by the noise in $\{\bar{Y}_i : i = 1, 2, \ldots, n\}$. Whereas their approach shrinks every $\bar{Y}_i$ towards the grand average by the same shrinkage factor, our deconvolution estimator is a more principled approach that adjusts the estimated density in a more nuanced manner.

To make a more concrete example, consider a call center simulation in which the calls arrive at the center according to a Poisson process, whose mean arrival rate parameter Λ is fixed but unknown. See References [2, 15, 16, 23, 25] for further details and other modeling considerations relevant to call center simulation, including the situation in which parameters like Λ vary stochastically over time. In our setting of a fixed but unknown Λ, suppose we represent its uncertainty by assuming it follows some lognormal distribution with known parameters. In real call center simulations, there may be many different and more complex sources of outer-level uncertainty, such as parameters for arrival distributions for many different categories of calls, nonparametric arrival distribution uncertainty, uncertainties in many different service time distributions for different categories of calls and for different servers, random variables that represent initial states of many different queues, and so on. However, we use the lognormal parametric uncertainty model for simplicity and transparency, and because our approach treats more complex uncertainty sources exactly the same as simple parametric uncertainty.

The service time for the arrived call is assumed to follow an exponential distribution with known parameters. We also assume the capacity of the system is K, which means that the maximum number of calls waiting in queue or being served is K and an arriving call is lost if it sees the system is full, i.e., if the system already has K calls when the new call arrives. Each call is associated with a random profit that is assumed to follow a normal distribution, whose parameters are known. We

study the behavior of the system within a specified unit of time, during which time the arrival rate parameter Λ is assumed constant, and we consider the effect of Λ varying according to its underlying lognormal distribution. The particular measure of interest is the expected increase in profit over a specified time interval that results from adding one additional server to the call center, which is a function of the unknown arrival rate parameter Λ. Thus, the r.v. X that is of interest is the difference in the conditional expectation of the profit (conditional on Λ) between an $M/M/c/K$ queue (referred to as System 1) and an $M/M/(c+1)/K$ queue (referred to as System 2), where c is the number of servers. In the simulation, one outer-level replicate consists of generating one random realization Λ_i from its lognormal distribution. One inner-level replicate consists of generating call arrivals from their Poisson process with parameter Λ_i, generating service times for the calls from their exponential distribution, assigning a random profit to each accepted call from its normal distribution, and calculating the total profit gained by adding the server for the collection of calls that were accepted. We conducted paired experiments for System 1 and System 2, meaning that for the ith outer-level replicate, the two systems share the same arrival rate parameter Λ_i (except for the common value of Λ_i, the two systems were simulated independently). Using a common value of arrival rate helps to isolate the effect of adding one more server on the system performance. See Reference [23] for further discussion of common random numbers in call center simulations. After obtaining the total profit for the two systems on each inner-level replicate j, the profit difference Y_{ij} between the two systems were then calculated. Note that given Λ_i, the randomness in Y_{ij} from the outer level is due to the random Poisson arrivals, as well as the random service times and random profit.

More specifically, for the jth inner-level replicate of the ith outer-level replicate, denote the total numbers of accepted calls within the specified time interval as N_{ij}^1 and N_{ij}^2 for System 1 and System 2, respectively. Each accepted call is associated with a random profit $P_{ij,k}^r, k = 1, 2, \ldots, N_{ij}^r; r = 1, 2$, and the total profit gained by System r within the specified time interval is $\sum_{k=1}^{N_{ij}^r} P_{ij,k}^r$. Therefore, the r.v. Y_{ij}, which is the observed increase in profit by adding one more server in the call center system, is

$$Y_{ij} = \sum_{k=1}^{N_{ij}^2} P_{ij,k}^2 - \sum_{k=1}^{N_{ij}^1} P_{ij,k}^1.$$

We are interested in estimating the pdf of the expected increase in profit given the random arrival rate Λ, i.e., the pdf of $X = \mathrm{E}[Y_{ij}|\Lambda]$. The sample average $\bar{Y}_i = m^{-1} \sum_{j=1}^m Y_{ij}$ can be viewed as a noisy estimate of $X_i = \mathrm{E}[Y_{ij}|\Lambda_i]$.

For the parameterization of this example, we take $\Lambda \sim \mathrm{lognorm}(1.68, 0.3)$, which has mean 5.61 and standard deviation 1.72. The service time follows an exponential distribution with a mean of 1. The distribution of the profit associated with each accepted call is chosen to be $N(1, 0.05)$. The number of servers is $c = 5$ for System 1 and $c+1 = 6$ for System 2, and the capacity for both systems is $K = 10$. For the purpose of having a closed-form analytical expression for f_X, the simulation is initiated in the steady state. (The closed-form expression for f_X is used only to evaluate the performance of the approach; we did not incorporate this knowledge of f_X into the estimation procedure.)

Figure 6 illustrates the typical noise level for one simulated dataset with the total computational budget being $M = 1.25 \times 10^6$ and the allocation being $m = 250$ and $n = 5{,}000$. The number of inner-level replicates m used in this example is larger than in the portfolio loss example, because the inherent inner-level noise in this example is much larger. Thus, even with this larger m, the naïve estimator is still substantially over-dispersed compared to the true f_X, as seen in the

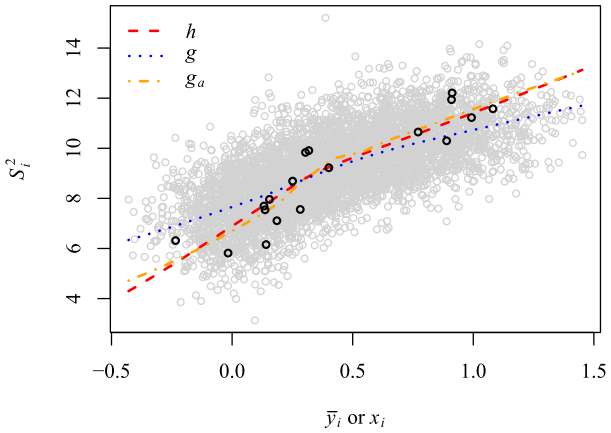


Fig. 6. For the call center example, a scatterplot of the sample variance S_i^2 against the sample mean $\bar{Y}_i$ for $i = 1, \ldots, n$, along with the "true" h function (the red dashed curve), the loess estimates g (the blue dotted curve) and g_a (the orange dot-dashed curve). The black open circles correspond to the replicates in the additional simulations that were used to estimate g_a.

histogram in Figure 7. Using too few inner-level replicates, the noise levels are so large that even the deconvolution or bias correction is challenging. Figure 6 is a scatterplot of S_i^2 vs. $\bar{Y}_i$ ($i = 1, 2, \ldots, n$) for each scenario ω_i. The "true" function h in this example is obtained by fitting a loess [6, 14] local linear smoother to a scatterplot of S_i^2 vs. $\bar{Y}_i$ from a separate data set with very large m and n values (which is not shown here and which was only used to obtain the ground truth for this example for assessment purposes). The estimated functions g and g_a (the latter fit to the additional simulation results, as described earlier) are also shown in Figure 6. It is again the case that g_a is much closer to the "true" function h compared to g, and the procedure for obtaining g_a only costs about 0.73% additional computational budget for this particular MC replicate. Across the MC replicates for this example, the additional cost for estimating g_a was generally between 0.5% and 0.9%.

Figure 7 compares the QP estimator with its alternatives, and it also demonstrates the enhanced performance achieved by including the support shape constraint in the QP method. Specifically, Figure 7 plots the histogram of the observations $\bar{y}_i$ for $i = 1, \ldots, n$, along with the true pdf f_X (the derivation of which is presented in Appendix C) and the four pdf estimators. The true pdf for the call center problem has an unusual shape due to the particular structure and parameterization of the problem, which makes the bias-correction challenging for this example. In spite of the challenging nature of this example, the QP estimator, which is denoted as "QP" and uses g_a and the two universally applicable shape constraints (integrate-to-one and nonnegativity), still performs much better than the naïve estimator at estimating f_X. In addition, the QP estimator is less over-dispersed than is the Steckley et al. estimator. We also plot the QP estimator with an additional constraint that the support of X is nonnegative (i.e., that $f_X(x) = 0$ for $x < 0$), which is denoted as "QPs" in Figure 7. Including the additional support constraint significantly improves the performance of the QP estimator. Specifically, QPs better captures the sharp feature at $x = 0$, which is difficult to accomplish using a conventional kernel-based method. It is reasonable to include the nonnegative support constraint for this particular call center example, because we know that given the same arrival rate and call center capacity, adding a server to the system can only increase the expected number of customers served, and hence can only increase the expected profit (ignoring the cost of the server).

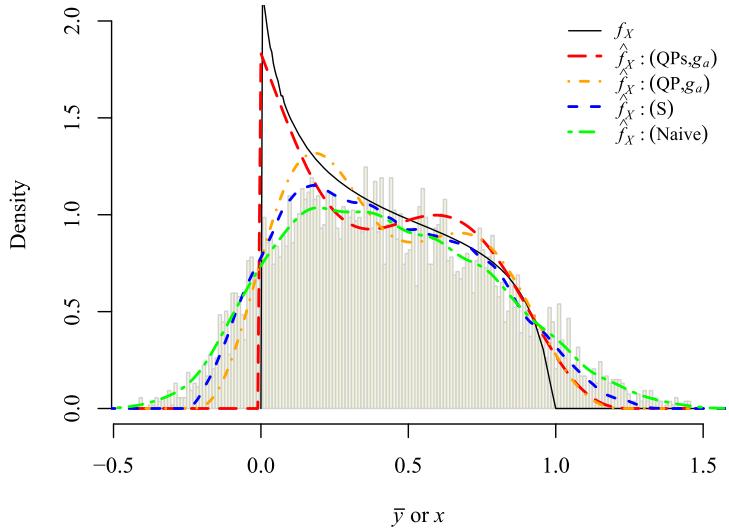


Fig. 7. Histogram of the observations $\bar{Y}_i$ ($i = 1, \ldots, n$) for the call center example, along with the true pdf f_X (black solid curve) and two QP estimators using g_a (the red long-dashed curve and the orange dot-dashed curve). The two QP estimators both incorporate the integrate-to-one and the nonnegativity constraints, while the one denoted as QPs (red long-dashed curve) includes the additional support constraint that $X \geq 0$. The blue dashed curve is the Steckley et al. estimator, and the green two-dashed curve is the naïve estimator (the kernel-smoothed histogram).

3.2 Portfolio Loss Example with Adaptive Allocation

In some financial applications, the inner-level allocation $m = m(\omega_i) = m(X(\omega_i))$ is determined by an adaptive procedure and depends on the individual outer-level scenario ω_i References [5, 12]. As discussed in Section 2.3, our QP estimation approach remains applicable when the nested simulation utilizes adaptive allocation as long as the number of inner-level replicates does not depend on ω_i other than via $X_i = X(\omega_i)$. In this section, we examine the performance of the QP estimator when an adaptive allocation is used in the portfolio loss example introduced in Section 3.1.1. An adaptive allocation is useful mainly when, instead of desiring the entire pdf f_X, one is primarily interested in $Pr[X \leq q]$ for one particular threshold q, where q may be prespecified or may be a specified quantile.

As was demonstrated in Reference [5], the number of inner-level replicates in the adaptive allocation should be larger, perhaps orders of magnitude larger, in the vicinity of the threshold q. Here, we assume the rule for choosing m_i as a function of the portfolio loss $X(\omega_i)$ for the ith scenario is given, and we focus on the behavior of the different pdf estimators under the adaptive allocation. Since X is not known in practice, one would have to choose m_i as a function of some surrogate like $\bar{Y}_i$, as was done in References [5, 12]. Figure 8 plots our rule for choosing $m(x)$ as a function of x, which is based on the functional relation suggested in Reference [5] for a specified loss threshold $q = 0.721$, which corresponds to the 0.90 quantile of f_X. In practice, the 0.90 quantile $q = 0.721$ would not be known in advance. Regardless, we still use $Pr[X \leq 0.721]$ as a convenient basis for comparison in this example. Figure 8(a) is the theoretical rule, while Figure 8(b) is the actual rule that rounds $m(x)$ to the nearest integer. We use $n = 4,096$ as the fixed number of outer-level replicates, and the total computational budget for this example is 80,000. The functions h, g, and g_a for one typical MC replicate are plotted in Figure 9. The corresponding conditional error variance functions for $\bar{Y}$ plotted in Figure 9(b) are obtained by dividing the functions in Figure 9(a) by the

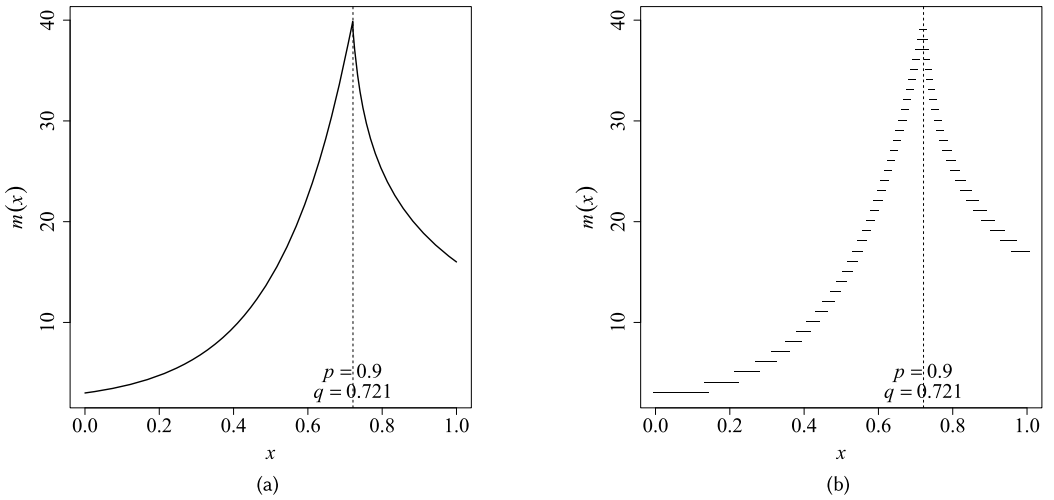


Fig. 8. (a): The theoretical number of inner-level replicates $m(x)$ as a function of x for each scenario for the adaptive allocation of the portfolio loss example. (b): The actual integer number of inner-level replicates used in the simulation. The vertical dashed line corresponds to the 0.90 quantile at $q = 0.721$.

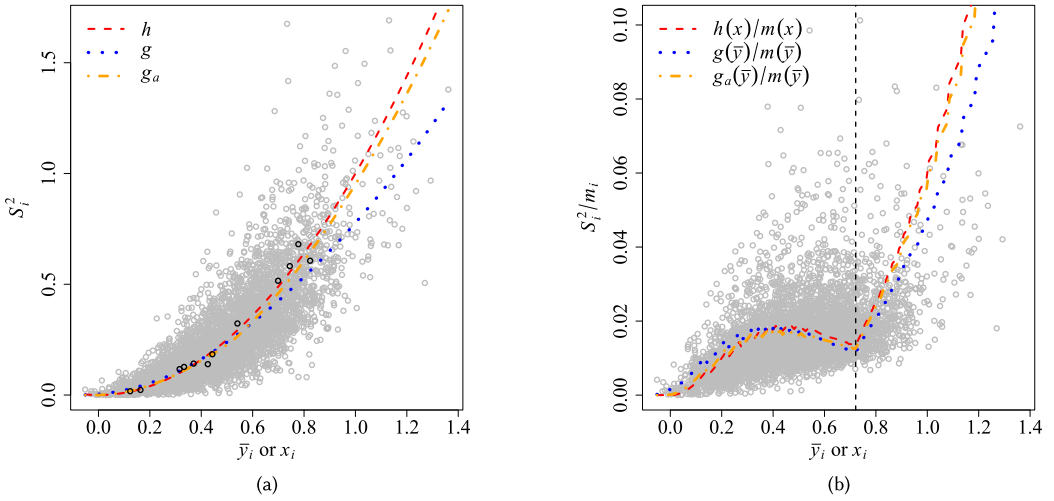


Fig. 9. (a): For the adaptive allocation of the portfolio loss example, the theoretical h function (red dashed curve) and the estimated functions with (g_a: orange dot-dashed curve) and without (g: blue dotted curve) additional simulations. The black open circles correspond to the additional replicates from which g_a was obtained. (b): The functions in Figure 9(a) divided by the $m(x)$ function shown in Figure 8(b) serve as estimates of the variance function in Figure 9(b). The vertical dashed line corresponds to the 0.90 quantile $q = 0.721$.

corresponding number of inner-level replicates shown in Figure 8(b). Notice that the conditional error variance functions dip around the threshold $q = 0.721$ due to the increased number of the inner-level replicates in that vicinity, which is intuitively appealing if the goal is to estimate the 0.90 quantile. The histogram of the $\bar{Y}$ observations for this MC replicate is displayed in Figure 10, and the histogram mass to the left of the vertical line at the true 0.90 quantile $q = 0.721$ is only 0.854. That is, the estimate of $Pr[X \leq 0.721]$ using the naïve histogram estimator is only 0.854,

Table 2. Comparison of the Averages and MSEs of the Estimated $Pr[X \leq q]$ for Three Methods across 200 MC Replicates for the Adaptive Allocation Version of the Portfolio Loss Example

p	Naïve		Steckley		QP g_a	
	$Ave(\hat{p})$	$MSE(\hat{p})$	$Ave(\hat{p})$	$MSE(\hat{p})$	$Ave(\hat{p})$	$MSE(\hat{p})$
0.8	0.771	8.74e-04 (2.95e-05)	0.800	8.16e-05 (8.71e-06)	0.808	9.16e-05 (8.59e-06)
0.9	0.850	2.51e-03 (3.42e-05)	0.883	3.29e-04 (1.74e-05)	0.902	8.56e-05 (7.39e-06)
0.95	0.898	2.70e-03 (3.33e-05)	0.931	4.16e-04 (1.82e-05)	0.944	7.47e-05 (6.88e-06)
0.99	0.953	1.41e-03 (1.79e-05)	0.978	1.62e-04 (8.01e-06)	0.985	3.53e-05 (2.80e-06)

Note: The Quantiles q Correspond to Probability Levels $\{0.80, 0.90, 0.95, 0.99\}$ Standard errors of the MSEs are shown in parentheses.

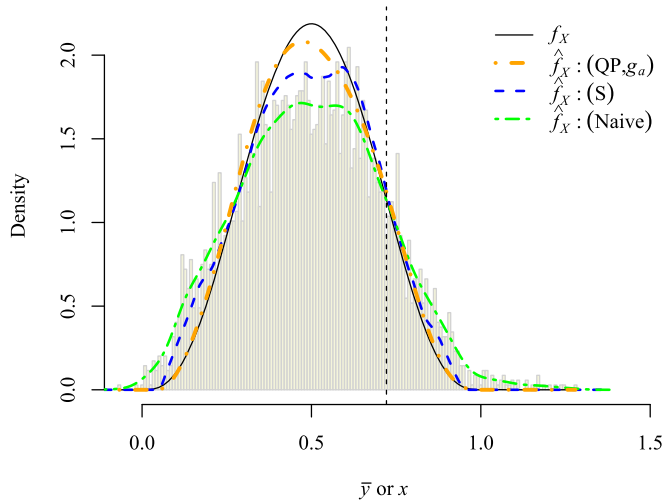


Fig. 10. For the adaptive allocation of the portfolio loss example, the histogram (naïve estimator) of the loss density together with three pdf estimators. The orange dotted dashed curve is the QP estimator with g_a, the blue dashed curve is the Steckley et al. estimator, and the green dashed curve is the kernel smoothing estimator. The vertical dashed line corresponds to the 0.90 quantile $q = 0.721$.

compared to the true probability 0.90. In contrast, the estimate of $Pr[X \leq 0.721]$ using the QP estimator with g_a is 0.899, which is very close to the true probability 0.90.

The preceding results are for one typical MC replicate. We repeated the analysis for 200 MC replicates, the results of which are summarized in Table 2. On each MC replicate, the two-level simulation was repeated, and the three pdf estimators (naïve, Steckley et al., and QP) were re-calculated. For each of the three pdf estimators, the corresponding estimate $\hat{p}$ of the probability $Pr[X \leq q]$ for four separate tail quantiles q corresponding to probabilities $p \in \{0.8, 0.9, 0.95, 0.99\}$ were calculated. Table 2 shows the average $\hat{p}$ values across the 200 replicates, along with their MSEs, from which we see that using bias-corrected estimators (the Steckley et al. estimator and the QP estimator) improves the naïve estimator by an order of magnitude. Furthermore, the QP estimator outperforms the Steckley et al. estimator (and the naïve estimator) in terms of both bias and MSE.

4 CONCLUSIONS

In this article, we have developed and analyzed a density deconvolution approach for obtaining a bias-corrected estimator of the density of the conditional expectation from a nested simulation.

We have derived the deconvolution relationships, which are valid for either uniform (constant $m_i = m$) or adaptive (variable m_i) allocations, as long as the number m_i of inner-level replicates is chosen to depend on the outer-level scenario ω_i only via $X_i = X(\omega_i)$.

Our deconvolution estimator requires an estimate of the conditional variance function $h(x)$, which is challenging due to the "errors-in-predictors" phenomenon that results from $\bar{Y}_i$ differing from X_i due to finite m_i. Our solution is to conduct additional inner-level replicates for just a handful of outer-level replicates. This was to ensure that we obtain a limited number of higher-quality $\bar{Y}_i$ observations that are closer to the true X_i and also to ensure that the sample variances for the inner-level replicates are close to the true conditional variances. We fit a local linear loess smoother to these fewer but higher-quality data points to get the conditional error variance function. We found this approach to provide a more accurate error variance function estimator, compared to simply fitting a loess smoother to all the lower-quality data points (i.e., the $\bar{Y}_i$ observations obtained in the main nested simulation using the fewer number of inner-level replicates) (see Figures 2 and 6).

The examples in Section 3.1 have demonstrated that the deconvolution estimator with only the universally applicable integrate-to-one and nonnegative constraints outperforms the naïve estimator and the Steckley et al. estimator, and the difference margin is largest when the number of inner-level replicates is small. What is particularly relevant is that the deconvolution estimator performs better than the alternative estimators across the range of quantiles when each estimator uses its optimal allocation for a fixed computational budget (see Figure 3, Figure 4, and Table 1), and its performance does not degrade as much when m differs from its optimal value (see Figsures 1 and 4). This form of robustness to choice of m in the allocation between inner- and outer-level replicates is important, because it is difficult to determine the optimal m in practice, and there is no single optimal m if multiple features of f_X are of interest (e.g., its mean, median, standard devaition, multiple upper quantiles, multiple lower quantiles). Another advantage of the deconvolution estimator is that a number of relevant shape constraints on the pdf can be easily incorporated, which further improves its performance. This was illustrated via the call center example, in which including an additional $X \geq 0$ support shape constraint on the pdf f_X further improved the deconvolution estimator by a substantial amount (Figure 7). Finally, the deconvolution method performed well not only in the uniform allocation setting, but also in the adaptive allocation nested simulation setting, as demonstrated in Section 3.2.

APPENDIX

A DERIVATION OF EQUATION (12)

We derive Equation (12) based on the second-order Taylor expansion approximation of Equation (11) to establish the predominant dependence of $f_{\bar{Y}}$ on f_X and on the conditional error variance function $v(\cdot)$. From Equation (11),

$$\begin{aligned} f_{\bar{Y}}(y) &\cong \int_{-\infty}^{\infty} \left\{ f_{X,Z}(y,z) - f'_{X,Z}(y,z)z + \frac{1}{2} f''_{X,Z}(y,z)z^2 \right\} \mathrm{d}z \\ &= \int_{-\infty}^{\infty} f_{X,Z}(y,z)\mathrm{d}z - \int_{-\infty}^{\infty} f'_{X,Z}(y,z)z\mathrm{d}z + \int_{-\infty}^{\infty} f''_{X,Z}(y,z)z^2/2\mathrm{d}z. \end{aligned}$$

Assuming regularity conditions [26] that allow us to exchange the order of the differentiation and integration operations and since $\mathrm{E}[Z|X = y] = 0$, we have

$$f_{\bar{Y}}(y) = f_X(y) - \frac{\partial}{\partial y}\int_{-\infty}^{\infty} f_{X,Z}(y,z)z\mathrm{d}z + \frac{1}{2}\frac{\partial^2}{\partial y^2}\int_{-\infty}^{\infty} f_{X,Z}(y,z)z^2\mathrm{d}z \tag{17}$$

$$= f_X(y) - \frac{\partial}{\partial y} f_X(y)\int_{-\infty}^{\infty} f_{Z|X}(z|y)z\mathrm{d}z + \frac{1}{2}\frac{\partial^2}{\partial y^2} f_X(y)\int_{-\infty}^{\infty} f_{Z|X}(z|y)z^2\mathrm{d}z \tag{18}$$

$$= f_X(y) - \frac{\partial}{\partial y}\Big\{f_X(y)\mathrm{E}[Z|X=y]\Big\} + \frac{1}{2}\frac{\partial^2}{\partial y^2}\Big\{f_X(y)\mathrm{E}[Z^2|X=y]\Big\} \tag{19}$$

$$= f_X(y) - \frac{\partial}{\partial y}\Big\{f_X(y)\mathrm{E}[Z|X=y]\Big\} + \frac{1}{2}\frac{\partial^2}{\partial y^2}\Big\{f_X(y)v(y)\Big\} \tag{20}$$

$$= f_X(y) + \frac{1}{2}v(y)f_X''(y) + v'(y)f_X'(y) + \frac{1}{2}v''(y)f_X(y). \tag{21}$$

B DERIVATION OF SIGNAL-TO-NOISE RATIO IN SECTION (3.1.1)

The derivation of the signal-to-noise ratio mentioned in the portfolio loss example in Section (3.1.1) follows from the familiar variance decomposition formula and from our specific parametrization of the example. Specifically, since the r.v. $Z|X = x$ is defined as $m^{-1}\sum_{j=1}^{m}\epsilon_j|X=x$, under the assumptions that $\epsilon_j|X = x$ are i.i.d. for $j = 1,\ldots,m$ and $\epsilon_j|X = x \sim Gamma(4, 2/x) - 2x$, the conditional expectation and conditional variance of $Z|X = x$ are 0 and x^2/m, respectively. The variance decomposition formula becomes

$$\mathrm{Var}[Z] = \mathrm{E}\Big[\mathrm{Var}[Z|X]\Big] + \mathrm{Var}\Big[\mathrm{E}[Z|X]\Big] \tag{22}$$

$$= \mathrm{E}\left[\frac{X^2}{m}\right] + \mathrm{Var}[0] \tag{23}$$

$$= \frac{1}{m}\Big[\mathrm{Var}[X] + \mathrm{E}^2[X]\Big]. \tag{24}$$

Hence, the signal-to-noise ratio is

$$\frac{\mathrm{Var}[X]}{\mathrm{Var}[Z]} = \frac{m\mathrm{Var}[X]}{\mathrm{Var}[X] + \mathrm{E}^2[X]}.$$

When $X \sim Beta(4,4)$, we have $\mathrm{Var}[X] = 1/36$, $\mathrm{E}[X] = 1/2$, and $\mathrm{Var}[X]/\mathrm{Var}[Z] = m/10$.

C DERIVATION OF THE THEORETICAL f_X IN THE CALL CENTER EXAMPLE

In this section, we derive the analytical expression for the theoretical pdf f_X in the call center example. For an $M/M/c/K$ queue, define $\rho = \lambda/\mu$ to be the ratio of the arrival rate λ over the service rate μ (here, μ is known), and define $a = \rho/c$. The number of calls in an $M/M/c/K$ system has a steady state distribution whose probability mass function is [31, 32]

$$p_n = Pr(\#\text{ of calls in the system} = n|\text{ arrival rate} = \lambda) = \begin{cases} \frac{\rho^n}{n!}p_0: & 0 \le n \le c \\ \frac{\rho^n c^{c-n}}{c!}p_0: & c \le n \le K \end{cases},$$

where

$$p_0 = Pr(\text{no call in the system}|\text{ arrival rate} = \lambda) = \begin{cases} \left[\frac{\rho^c}{c!}\cdot\frac{1-a^{K-c+1}}{1-a} + \sum_{n=0}^{c-1}\frac{\rho^n}{n!}\right]^{-1}: & a \ne 1 \\ \left[\frac{\rho^c}{c!}(K-c+1) + \sum_{n=0}^{c-1}\frac{\rho^n}{n!}\right]^{-1}: & a = 1 \end{cases}.$$

Then, we define

$$p_{-K} = Pr(\text{an arriving call is accepted}|\ \text{arrival rate} = \lambda) = 1 - p_K = 1 - \frac{\rho^K c^{c-K}}{c!} p_0.$$

Hence, by Poisson splitting, the number of accepted calls in one unit of time is $Pois(\lambda p_{-K})$, given that the arrival rate in this unit of time is λ.

Let N_r denote the number of accepted calls for system $r, r = 1, 2$ (System 1 is $M/M/c/K$; System 2 is $M/M/(c+1)/K$). Then, within a unit of time with arrival rate λ, we have $N_r|\Lambda = \lambda \sim Pois(\lambda p^r_{-K})$, where p^r_{-K} is the probability of an arriving call being accepted given an arrival rate λ for System r ($r = 1, 2$). Denote the difference in the total profit gained by the two systems as Y, whose conditional expectation given the unknown arrival rate Λ is the expected profit improvement that results from adding one more server, which we denote by X. Thus, we have

$$X = \mathrm{E}[Y|\Lambda] = \mathrm{E}\left[\left(\sum_{k=1}^{N_2} P_{2,k} - \sum_{k=1}^{N_1} P_{1,k}\right)\Big|\Lambda\right] = \mathrm{E}[P_{1,1}] \cdot \mathrm{E}[(N_2 - N_1)|\Lambda], \tag{25}$$

where the $P_{r,k}$ denotes the profit gained by the kth accepted call in System r ($r = 1, 2$). The second equality in Equation (25) holds due to the fact that all random profits are i.i.d. Since the distribution of the profit associated with each accepted call is chosen to be $N(1, 0.05)$, in which case $\mathrm{E}[P_{r,k}] = 1$ for all r and k, Equation (25) becomes $X = \mathrm{E}[(N_2 - N_1)|\Lambda]$. If we define $\mathrm{E}[(N_2 - N_1)|\Lambda = \lambda] = \lambda p^2_{-K} - \lambda p^1_{-K} \triangleq t(\lambda)$, then $X = t(\Lambda)$ with pdf

$$f_X(x) = f_{t(\Lambda)}(x) = f_\Lambda\left(t^{-1}(x)\right)\left|\frac{\mathrm{d}}{\mathrm{d}x} t^{-1}(x)\right|.$$

Remark C1: Note that $t(\lambda) = \lambda p^2_{-K} - \lambda p^1_{-K}$, where $p^1_{-K} = 1 - \frac{\rho^K c^{c-K}}{c!} \cdot [1 + \rho + \rho^2/2! + \cdots + \frac{\rho^c}{c!}\frac{1-a^{K-c+1}}{1-a}]^{-1} \approx 1 - \frac{\rho^K c^{c-K}}{c!} \cdot (1-\rho)$ and $p^2_{-K} \approx 1 - \frac{\rho^K (c+1)^{c-K}}{c!} \cdot (1-\rho)$. Therefore,

$$t(\lambda) = \lambda\left(1 - \frac{\lambda}{\mu}\right)\frac{\lambda}{\mu}^K \frac{c^{c-K} - (c+1)^{c-K}}{c!} \propto \lambda^{K+1},$$

from which it follows that

$$\frac{\mathrm{d}}{\mathrm{d}x} t^{-1}(0) \propto \frac{\mathrm{d}}{\mathrm{d}x} \sqrt[K+1]{x}\Big|_{x=0} \propto x^{-\frac{K}{K+1}}\Big|_{x=0} = \infty.$$

Since $f_X(x) = f_\Lambda(t^{-1}(x))|\frac{\mathrm{d}}{\mathrm{d}x} t^{-1}(x)|$, and the derivative of $t^{-1}(x)$ at $x = 0$ is infinite by the above argument, the pdf of X at 0 is infinite, i.e., $f_X(0) = \infty$. This is not apparent in Figure 7, since the vertical axis was truncated.

D CONSISTENCY AND RATE OF CONVERGENCE

In this section, we present a proof that QP deconvolution estimator is weakly consistent and a heuristic argument that its convergence rate is the same as the rate of the jackknife estimator in Steckley et al. [28]. The many challenging steps and technical details needed to convert the heuristic argument to a rigorous proof would require a new article.

D.1 Consistency

In this section, we address non-parametric consistency of the estimator defined in Equation (15). Since that equation only defines the estimate at the K_n discretization points, $\widehat{f}_{X,n}(x_{n,i})$, $i = 1, \ldots, K_n$, we must extend the solution to a function on $\mathbb{R}$. We do that by taking $\widehat{f}_{X,n}$ to be a histogram-like estimate: the function, constant on $[x_{n,i} - \delta_n/2, x_{n,i} + \delta_n/2), i = 1, \ldots, K_n$ with values on those intervals given by the solution of Equation (15), and zero elsewhere. Denoting

the integral transform in Equation (8) by $\mathcal{K}f(y) := \int_{-\infty}^{\infty} f_{Z|X}(y-x|x)f(x)\,\mathrm{d}x$, we roughly follow the approach in Mendelsohn and Rice [24]: We first show in Theorem 2 that under certain conditions, $\mathcal{K}\widehat{f}_{X,n}$ is strongly $\mathcal{L}_2$-consistent for $f_{\bar{Y}} = \mathcal{K}f_X$. Then, in Theorem 3, we show that as a consequence of Theorem 2, the corresponding CDFs $\widehat{F}_{X,n}$ converge weakly to F_X with probability 1 (weak convergence here meaning pointwise convergence at every continuity point of F_X; in this case every point, since F_X is assumed to have a density). Finally, in Lemma 2, we show that if f_X has four $\mathcal{L}_2$ derivatives, and $f_{Z|X}$ is Gaussian with variance function $v(x)$, then under some conditions on $v(x)$, the assumptions of Theorem 2 are satisfied.

In this section, matrices and vectors are written in boldface, and if f denotes a compactly supported, piecewise constant function, then $\mathbf{f}$ denotes the vector of the values it attains. Operators on function spaces are denoted by calligraphic letters, so $\mathcal{C}f$ maps the function f to a function, while $\mathbf{Cf}$ maps the vector $\mathbf{f}$ to a vector. The norm on the function space $\mathcal{L}_2(\mathbb{R})$ is denoted by $\|\cdot\|_{\mathcal{L}_2}$ and is defined by $\|f\|_{\mathcal{L}_2}^2 = \int_{-\infty}^{\infty} f(x)^2\,dx$. The conventional 2-norm on $\mathbb{R}^K$ is denoted by $\|\cdot\|_{K,2}$. Note that if $\hat{f}$ is a random function, or $\hat{\mathbf{f}}$ a random vector in $\mathbb{R}^K$, then $\|\hat{f}\|_{\mathcal{L}_2}$ and $\|\hat{\mathbf{f}}\|_{K,2}$ are both $\mathbb{R}$-valued random variables.

Theorem 2. *For each integer n, let the approximation subspace $\mathscr{A}_n$, with approximation interval $A_n := [x_{n,1} - \delta_n/2, x_{n,K_n} + \delta_n/2)$, consist of non-negative functions integrating to unity, with support contained in A_n, constant on $[x_{n,i} - \delta_n/2, x_{n,i} + \delta_n/2)$, and uniformly bounded by some $B > 0$. Assume the following:*

(1) There is a sequence $\{a_n\}$, each $a_i \in \mathscr{A}_i$, with the property that $\|a_n - f_X\|_{\mathcal{L}_2} \to 0$ and $\delta_n\|\boldsymbol{D}_2\boldsymbol{a}_n\|_{K_n,2}^2 = O(1)$.

(2) There is a sequence of approximators $\{\widehat{f}_{\bar{Y},n}\}$ of $f_{\bar{Y}}$ with $\widehat{f}_{\bar{Y},i} \in \mathscr{A}_i$ s.t. $\|\widehat{f}_{\bar{Y},n} - f_{\bar{Y}}\|_{\mathcal{L}_2} \overset{a.s.}{\to} 0$.

(3) The functions $\widehat{f}_{X,n}$ and a_n have support contained in $S_n = [a_n, b_n]$, and the maps $y \mapsto \mathbf{1}_{A_n^c}(y)\sup_{x\in S_n} f_{Z|X}(y-x|x)^2$ are bounded by some integrable function.

(4) The operator $\mathcal{K} : \mathcal{L}_2(\mathbb{R}) \to \mathcal{L}_2(\mathbb{R})$ is bounded.

(5) For each n, our estimate is the unique $\widehat{f}_{X,n} \in \mathscr{A}_n$ minimizing $\|\widehat{\boldsymbol{f}}_{\bar{Y},n} - \boldsymbol{C}_n\widehat{\boldsymbol{f}}_{X,n}\|_{K_n,2}^2 + \lambda_n\|\boldsymbol{D}_2\widehat{\boldsymbol{f}}_{X,n}\|_{K_n,2}^2$, where $\widehat{\boldsymbol{f}}_{\bar{Y},n}$ and $\boldsymbol{C}_n\widehat{\boldsymbol{f}}_{X,n}$ are vectors of values of $\widehat{f}_{\bar{Y},n}$ and $C_n\widehat{f}_{X,n}$ and $\boldsymbol{D}_2$ a second-differencing operator.

(6) The map $(x,y) \mapsto f_{Z|X}(y-x|x)$ is uniformly continuous in (x,y) with modulus of continuity $\omega(\varepsilon)$.

(7) $\lambda_n \to 0$ and $\omega(\delta_n)(b_n - a_n)\sqrt{K_n\delta_n} \to 0$.

Then,

$$\|\mathcal{K}f_X - \mathcal{K}\widehat{f}_{X,n}\|_{\mathcal{L}_2} \overset{a.s.}{\to} 0. \tag{26}$$

Proof of Theorem 2. The formulation in Equation (15) uses two layers of approximation. First, only a finite interval is addressed in the objective, while $\mathcal{K}\widehat{f}_{X,n}$ may have support on all of $\mathbb{R}$; second, the integral transform $\mathcal{K}$ is approximated by a discrete approximation. If we are given $K_n, \delta_n, \{x_{n,i}\}$, with approximation interval $A_n := [x_{n,1} - \delta_n/2, x_{n,K_n} + \delta_n/2)$, with $\mathbf{C}_n$ the convolution matrix $\mathbf{C}$ in Equation (14), and the $\tilde{x}_i \in [x_{n,i} - \delta_n/2, x_{n,i} + \delta_n/2)$ satisfying $\delta_n f(\tilde{x}_i)f_{Z|X}(y-\tilde{x}_i|\tilde{x}_i) = \int_{x_{n,i}-\delta_n/2}^{x_{n,i}+\delta_n/2} f(x)f_{Z|X}(y-x|x)\mathrm{d}x$, then we can define the following mappings to address each of those approximations:

$$C_n f(y) = \begin{cases} \delta_n[\mathbf{C}_n\mathbf{f}]_j = \delta_n\sum_{i=1}^{K_n} f(x_{n,i})f_{Z|X}(x_{n,j} - x_{n,i}|x_{n,i}) & \text{if } y \in [x_{n,j} - \delta/2, x_{n,j} + \delta_n/2) \\ 0 & \text{if } y \notin A_n, \end{cases} \tag{27}$$

$$\mathcal{R}_n f(y) = \begin{cases} \delta_n \sum_{i=1}^{K_n} [f(\tilde{x}_i) f_{Z|X}(y - \tilde{x}_i|\tilde{x}_i) - f(x_{n,i}) f_{Z|X}(x_{n,j} - x_{n,i}|x_{n,i})] & \text{if } y \in [x_{n,j} - \delta/2, x_{n,j} + \delta_n/2) \\ 0 & \text{if } y \notin A_n, \end{cases} \tag{28}$$

$$\mathcal{T}_n f(y) = \begin{cases} 0 & \text{if } y \in A_n \\ \int_{-\infty}^{\infty} f(x) f_{Z|X}(y - x|x) \mathrm{d}x & \text{if } y \notin A_n, \end{cases} \tag{29}$$

representing the approximation of the convolution, the remainder from that approximation, and the remainder due to the ignored tail, respectively. We have the property that when f has support contained in A_n and is piecewise constant on $[x_{n,i} - \delta_n, x_{n,i} + \delta_n)$, then $\mathcal{K} f = C_n f + \mathcal{R}_n f + \mathcal{T}_n f$, since if $y \in [x_{n,j} - \delta_n, x_{n,j} + \delta_n) \subset A_n$,

$$\begin{aligned} \mathcal{K} f(y) &= \int_{A_n} f(x) f_{Z|X}(y - x|x) \mathrm{d}x \\ \text{(a)} \quad &= \sum_{i=1}^{K_n} \int_{x_{n,i} - \delta_n/2}^{x_{n,i} + \delta_n/2} f(x) f_{Z|X}(y - x|x) \mathrm{d}x \\ \text{(b)} \quad &= \delta_n \sum_{i=1}^{K_n} f(\tilde{x}_i) f_{Z|X}(y - \tilde{x}_i|\tilde{x}_i) \\ \text{(c)} \quad &= \delta_n \sum_{i=1}^{K_n} f(x_{n,i}) f_{Z|X}(x_{n,j} - x_{n,i}|x_{n,i}) \\ &\quad + \delta_n \sum_{i=1}^{K_n} [f(\tilde{x}_i) f_{Z|X}(y - \tilde{x}_i|\tilde{x}_i) - f(x_{n,i}) f_{Z|X}(x_{n,j} - x_{n,i}|x_{n,i})] \\ &= C_n f(y) + \mathcal{R}_n f(y) \end{aligned} \tag{30}$$

In (a), we have re-written the integral as a sum over the discretization grid. In (b), we invoke the mean value theorem to get $\tilde{x}_i \in [x_{n,i} - \delta_n/2, x_{n,i} + \delta_n/2)$ s.t. $\int_{x_{n,i} - \delta_n/2}^{x_{n,i} + \delta_n/2} f(x) f_{Z|X}(y - x|x) \mathrm{d}x = \delta_n f(\tilde{x}_i) f_{Z|X}(y - \tilde{x}_i|\tilde{x}_i)$. In (c), we substitute the $\tilde{x}_i$'s and y's bin midpoints.

The foregoing is convenient, because if a function f is compactly supported and piecewise constant on a uniformly spaced partition—in our case, $C_n f$ is just such a function—then its $\mathcal{L}_2$-norm can be expressed in terms of the 2-norm of the vector of the values it takes on: If $\mathbf{f}$ is the vector with entries $\mathbf{f}_i = f(x_i)$, and the spacing is δ, then

$$\|f\|_{\mathcal{L}_2}^2 = \sum_{i=1}^{K} \delta f(x_i)^2 = \delta \|\mathbf{f}\|_{K,2}^2. \tag{31}$$

The objective in Equation (15) involves minimizing a penalized vector 2-norm $\|\widehat{\mathbf{f}}_{\tilde{Y},n} - \mathbf{C}\mathbf{f}\|_{K,2}^2$, and we are already thinking of $\widehat{f}_{\tilde{Y},n}$ as piecewise constant—a histogram—so it will be convenient to think of $\mathbf{C}_n\mathbf{f}$ as the values of a piecewise constant approximation $C_n f$ of $\mathcal{K} f$.

The following lemma will be useful. The lemma's proof is deferred to below the current proof.

Lemma 1. *Under the assumptions, the norms of the errors in our discrete approximation and support approximation converge to zero. Specifically,*

(i) Under Assumptions 3, 6, & 7, if $\{s_n\}$ is a sequence with $s_i \in \mathscr{A}_i$, then $\|\mathcal{R}_n s_n\|_{\mathcal{L}_2} \to 0$.
(ii) Under Assumption 3, if $\{s_n\}$ is a sequence of pdfs, then $\|\mathcal{T}_n s_n\|_{\mathcal{L}_2} \to 0$.

By Assumption 2, $\mathbb{P}(\|\widehat{f}_{\bar{Y},n} - f_{\bar{Y}}\|_{\mathcal{L}_2} \to 0) = 1$; assume that this event occurs. Since $f_{\bar{Y}} = \mathcal{K} f_X$, we want to show that $\|f_{\bar{Y}} - \mathcal{K}\widehat{f}_{X,n}\|_{\mathcal{L}_2}$ converges to zero. First, we bound it by an expression related to the objective in Equation (15) plus some approximation remainders.

$$\begin{aligned}
\|f_{\bar{Y}} - \mathcal{K}\widehat{f}_{X,n}\|_{\mathcal{L}_2} &= \|f_{\bar{Y}} - C_n\widehat{f}_{X,n} - \mathcal{R}_n\widehat{f}_{X,n} - \mathcal{T}_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} \\
&\le \|f_{\bar{Y}} - C_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} + \|\mathcal{R}_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} + \|\mathcal{T}_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} \\
&\le \|\widehat{f}_{\bar{Y},n} - C_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} + \|f_{\bar{Y}} - \widehat{f}_{\bar{Y},n}\|_{\mathcal{L}_2} + \|\mathcal{R}_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} + \|\mathcal{T}_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} \\
&= \sqrt{\delta_n}\|\widehat{\mathbf{f}}_{\bar{Y},n} - \mathbf{C}_n\hat{\mathbf{f}}_n\|_{K,2} + \|f_{\bar{Y}} - \widehat{f}_{\bar{Y},n}\|_{\mathcal{L}_2} + \|\mathcal{R}_n\widehat{f}_{X,n}\|_{\mathcal{L}_2} + \|\mathcal{T}_n\widehat{f}_{X,n}\|_{\mathcal{L}_2}.
\end{aligned} \tag{32}$$

The first term is what appears in the objective in Equation (15). As mentioned, the second term goes to zero by Assumption 2. The third and fourth terms go to zero by Lemma 1. To see that the first term goes to zero:

$$\begin{aligned}
\delta_n\|\widehat{\mathbf{f}}_{\bar{Y},n} - \mathbf{C}_n\widehat{\mathbf{f}}_{X,n}\|_{K,2}^2 &\le \delta_n[\|\widehat{\mathbf{f}}_{\bar{Y},n} - \mathbf{C}_n\widehat{\mathbf{f}}_{X,n}\|_{K,2}^2 + \lambda_n\|\mathbf{D}_2\widehat{\mathbf{f}}_{X,n}\|_{K,2}^2] \\
\text{(a)} &\le \delta_n[\|\widehat{\mathbf{f}}_{\bar{Y},n} - \mathbf{C}_n\mathbf{a}_n\|_{K,2}^2 + \lambda_n\|\mathbf{D}_2\mathbf{a}_n\|_{K,2}^2] \\
&= \delta_n\|\widehat{\mathbf{f}}_{\bar{Y},n} - \mathbf{C}_n\mathbf{a}_n\|_{K,2}^2 + \lambda_n\delta_n\|\mathbf{D}_2\mathbf{a}_n\|_{K,2}^2.
\end{aligned} \tag{33}$$

The first inequality follows from the positivity of $\lambda_n\|\mathbf{D}_2\widehat{\mathbf{f}}_{X,n}\|_2^2$. In (a), we use Assumption 5, and by Assumption 1, the second term goes to zero if $\lambda_n \to 0$. Finally, for the first term above,

$$\begin{aligned}
\sqrt{\delta_n}\|\widehat{\mathbf{f}}_{\bar{Y},n} - \mathbf{C}_n\mathbf{a}_n\|_2 &= \|\widehat{f}_{\bar{Y},n} - C_n a_n\|_{\mathcal{L}_2} \\
&\le \|f_{\bar{Y}} - \mathcal{K}a_n\|_{\mathcal{L}_2} + \|f_{\bar{Y}} - \widehat{f}_{\bar{Y},n}\|_{\mathcal{L}_2} + \|\mathcal{R}_n a_n\|_{\mathcal{L}_2} + \|\mathcal{T}_n a_n\|_{\mathcal{L}_2} \\
&= \|\mathcal{K}(f_X - a_n)\|_{\mathcal{L}_2} + \|f_{\bar{Y}} - \widehat{f}_{\bar{Y},n}\|_{\mathcal{L}_2} + \|\mathcal{R}_n a_n\|_{\mathcal{L}_2} + \|\mathcal{T}_n a_n\|_{\mathcal{L}_2}.
\end{aligned} \tag{34}$$

The first term converges to zero by Assumptions 4 and 1, the second by Assumption 2, and the third and fourth by Lemma 1.

Thus, we conclude that $\mathbb{P}(\|\mathcal{K}f_X - \mathcal{K}\widehat{f}_{X,n}\|_{\mathcal{L}_2} \to 0) = 1$, as needed. □

Proof of Lemma 1. For (i), let n be large enough that $A_n \supset [a_n, b_n]$ and suppose $y \in A_n$. We have, letting ℓ be the smallest index for which $x_\ell - \delta_n/2 > a_n$,

$$\begin{aligned}
\mathcal{R}_n s_n(y) &= \delta_n \sum_{i=1}^{K_n} s_n(x_{n,i})[f_{Z|X}(y - \tilde{x}_i|\tilde{x}_i) - f_{Z|X}(x_{n,j} - x_{n,i}|x_{n,i})] \\
\text{(a)} \qquad &\le B\delta_n \sum_{i=\ell}^{\ell+\lfloor(b_n-a_n)/\delta_n\rfloor} [f_{Z|X}(y - \tilde{x}_i|\tilde{x}_i) - f_{Z|X}(x_{n,j} - x_{n,i}|x_{n,i})] \\
\text{(b)} \qquad &\le B\delta_n \sum_{i=\ell}^{\ell+\lfloor(b_n-a_n)/\delta_n\rfloor} \omega(\delta_n) \le B(b_n - a_n)\omega(\delta_n),
\end{aligned} \tag{35}$$

where (a) is by the fact that $s_n < B$ and s_n is zero outside $[a_n, b_n]$, and (b) is by the uniform continuity of g with modulus $\omega(\varepsilon)$. Now, since $\mathcal{R}_n s_n(y) = 0$ for $y \notin A_n$,

$$\|\mathcal{R}_n s_n\|_{\mathcal{L}_2}^2 \le B^2(a_n - b_n)^2\omega(\delta_n)^2 K_n\delta_n, \tag{36}$$

so

$$\|\mathcal{R}_n s_n\|_{\mathcal{L}_2} \le B(a_n - b_n)\omega(\delta_n)\sqrt{K_n\delta_n}, \tag{37}$$

Assumption 7 gives the desired result.

Now, for (ii), observe that

$$\begin{aligned}\|\mathcal{T}_n s_n\|_{\mathcal{L}_2} &= \int_{A_n^c}\left[\int_{S_n} s_n(x) f_{Z|X}(y-x|x)\mathrm{d}x\right]^2 \mathrm{d}y \\ &\le \int_{A_n^c}\left[\left(\int s_n(x)\mathrm{d}x\right)\left(\sup_{x\in S_n} f_{Z|X}(y-x|x)\right)\right]^2 \mathrm{d}y \\ &= \int \mathbf{1}_{A_n^c}(y)\sup_{x\in S_n} f_{Z|X}(y-x|x)^2 \mathrm{d}y,\end{aligned} \tag{38}$$

invoking Hölder's inequality with $p = 1$, $q = \infty$. By the existence of a dominating integrable function from Assumption 3 and the fact that $\mathbf{1}_{A_n^c}(y)\sup_{x\in S_n} f_{Z|X}(y-x|x)^2$ converges pointwise to the zero function, the dominated convergence theorem yields that $\|\mathcal{T}_n s_n\|_{\mathcal{L}_2} \to 0$. □

Now, in the following, we upgrade result (26) about $\mathcal{K}\widehat{f}_{X,n}$ to convergence of $\widehat{F}_{X,n}$:

Theorem 3. *Let $\widehat{F}_{X,n}$ and F_X be the cdfs of $\widehat{f}_{X,n}$ and f_X, respectively. Suppose that f_X has the unique distribution for which $\mathcal{K}f_X = f_{\bar{Y}}$. Under Assumptions 1–7, $\mathbb{P}(\widehat{F}_{X,n} \xrightarrow{w} F_X) = 1$.*

Proof of Theorem 3. Under the assumptions, Theorem 2 yields that $\mathbb{P}(\|\mathcal{K}\widehat{f}_{X,n} - f_{\bar{Y}}\|_{\mathcal{L}_2} \to 0) = 1$. Assume that this event occurs. By Theorem 3.2.9 of Reference [9], if every subsequence of $\widehat{F}_{X,n}$ has a further subsequence converging to F_X, then $\widehat{F}_{X,n} \xrightarrow{w} F_X$. Let $\widehat{F}_{X,n_k}$ be a subsequence of $\widehat{F}_{X,n}$

By Helly's Selection Theorem (Theorem 3.2.6 of Reference [9]), there is a further subsequence such that $\widehat{F}_{X,n_{k_j}} \xrightarrow{w} \bar{F}$. Then, for any y, we have

$$\mathcal{K}\hat{f}_{X,n_{k_j}}(y) = \int f_{Z|X}(y-x|x)\mathrm{d}\widehat{F}_{X,n_{k_j}}(x) \to \int f_{Z|X}(y-x|x)\mathrm{d}\bar{F}(x) = \bar{f}_{\bar{Y}}(y). \tag{39}$$

But, since $\|\mathcal{K}\hat{f}_{X,n_{k_j}} - f_{\bar{Y}}\|_{\mathcal{L}_2} \to 0$, by Theorem 2, we have that $\|\bar{f}_{\bar{Y}} - f_{\bar{Y}}\|_{\mathcal{L}_2} = 0$. Hence $\bar{f}_{\bar{Y}} = f_{\bar{Y}}$ a.e., and by the assumption of uniqueness, $\bar{F} = F_X$. Thus, the conditions of Theorem 3.2.9 in Reference [9] are satisfied, and we conclude that $\widehat{F}_{X,n} \xrightarrow{w} F_X$. □

Lemma 2. *Assume that $f_X, f_X', \ldots, f_X^{(4)} \in \mathcal{L}_2$. Suppose that $A_n = [x_{n,i} - \delta_n/2, x_{n,K_n} + \delta_n/2]$ and $S_n = [a_n, b_n]$, with $S_n \subset A_n$. Suppose further that $a_n - (x_{n,1} - \delta_n/2)$ and $x_{n,K_n} + \delta_n - b_n$ are both increasing. Suppose that we have error densities*

$$f_{Z|X}(y-x|x) = \frac{1}{\sqrt{2\pi v(x)}}\exp -\frac{(y-x)^2}{2v(x)},$$

and that there are $m, M, m', M' \in \mathbb{R}$ s.t. $0 < m \le v(x) \le M$ and $m' \le v'(x) \le M'$ for all $x \in S$.

Then Assumptions 1–7 are satisfied.

Proof of Lemma 2.

Assumption 1. *One such sequence is given by*

$$a_n(x) = f_X(x_{n,i}) \qquad \text{for } x \in [x_{n,i} - \delta_n/2, x_{n,i} + \delta_n/2), i = 1, \ldots, K_n \tag{40}$$

and $a_n(x) = 0$ otherwise. Note that for $x \in [x_{n,i} - \delta_n/2, x_{n,i} + \delta_n/2)$,

$$|f_X(x) - a_n(x)|^2 = |f_X(x) - f_X(x_{n,i})|^2 \tag{41}$$

$$= \left|\int_x^{x_{n,i}} f'(t)\,dt\right|^2 \tag{42}$$

$$\leq \int_{x_{n,i}-\delta_n/2}^{x_{n,i}+\delta_n/2} f'(t)^2\,dt, \tag{43}$$

so

$$\|f_X - a_n\|_{\mathcal{L}_2}^2 = \int_{-\infty}^{\infty} |f_X(x) - a_n(x)|^2\,dx \tag{44}$$

$$= \sum_{i=1}^{K_n} \int_{x_{n,i}-\delta_n/2}^{x_{n,i}+\delta_n/2} |f_X(x) - a_n(x)|^2\,dx + \int_{A_n^c} f_X(x)^2\,dx \tag{45}$$

$$\leq \sum_{i=1}^{K_n} \int_{x_{n,i}-\delta_n/2}^{x_{n,i}+\delta_n/2} \int_{x_{n,i}-\delta_n/2}^{x_{n,i}+\delta_n/2} f'(t)^2\,dt\,dx + \int_{A_n^c} f_X(x)^2\,dx \tag{46}$$

$$= \delta_n \int_{A_n} f'(t)^2\,dt + \int_{A_n^c} f_X(x)^2\,dx \tag{47}$$

$$\leq \delta_n \|f_X'\|_{\mathcal{L}_2}^2 + \int_{A_n^c} f_X(x)^2\,dx, \tag{48}$$

each of which converges to zero as $n \to \infty$. *Now, let*

$$\mathcal{D}_2 a_n(x) = \begin{cases} \frac{a_n(x_{n,i}+\delta_n) - 2a_n(x_{n,i}) + a_n(x_{n,i}-\delta_n)}{\delta_n^2} & \text{if } x \in [x_{n,i} - \delta_n, x_{n,i} + \delta_n), i = 1, \ldots, K_n \\ 0 & \text{o.w.} \end{cases} \tag{49}$$

so by definition of a_n,

$$\mathcal{D}_2 a_n(x) = \begin{cases} \frac{f_X(x_{n,i}+\delta_n) - 2f_X(x_{n,i}) + f_X(x_{n,i}-\delta_n)}{\delta_n^2} & \text{if } x \in [x_{n,i} - \delta_n, x_{n,i} + \delta_n), i = 1, \ldots, K_n \\ 0 & \text{o.w.} \end{cases} \tag{50}$$

By the second representation above (see Reference [7]),

$$\mathcal{D}_2 a_n(x) = f_X''(x) - \frac{\delta_n^2}{12} f_X^{(4)}(\xi_x), \tag{51}$$

with $\xi_x \in [x_{n,i} - \delta_n, x_{n,i} + \delta_n)$ *if* $x \in [x_{n,i} - \delta_n, x_{n,i} + \delta_n)$, *and an argument similar to the above yields that* $\|\mathcal{D}_2 a_n\|_{\mathcal{L}_2}^2 = \delta_n \|\boldsymbol{D}_2 \boldsymbol{a}_n\|_{K_n,2}^2 = O(1)$.

Assumption 2. *Theorem 2 of Reference [37] gives that if* $f_{\bar{Y}} \in \mathcal{L}_2$, $x_{n,1} \to -\infty$, $x_{n,K} \to \infty$, $\delta_n \to 0$, *and* $n\delta_n \to \infty$, *then* $\|\widehat{f}_{\bar{Y},n} - f_{\bar{Y}}\|_{\mathcal{L}_2} \overset{a.s.}{\to} 0$.

Assumption 3. *If* $y < x_{n,1} - \delta_n$, *then*

$$\sup_{x \in S_n} \frac{1}{2\pi v(x)} \exp -\frac{(y-x)^2}{v(x)} \leq \frac{1}{2\pi m} \exp -\frac{(y-a_1)^2}{M} \tag{52}$$

and similarly if $y > x_{n,K_n} + \delta_n$, *then*

$$\sup_{x \in S_n} \frac{1}{2\pi v(x)} \exp -\frac{(y-x)^2}{v(x)} \leq \frac{1}{2\pi m} \exp -\frac{(y-b_1)^2}{M}, \tag{53}$$

each of which are integrable, so a satisfactory integrable function is found by taking the former for $y < a_1$, *the latter for* $y > b_1$, *and some constant value for* $y \in [a_1, b_1]$.

Assumption 4. *By Reference [19], Chapter 16, Theorem 2,* $\mathcal{K} : \mathcal{L}_2(\mathbb{R}) \to \mathcal{L}_2(\mathbb{R})$ *is bounded if*

$$\sup_x \int f_{Z|X}(y - x|x)\,dy < \infty \qquad \text{and} \qquad \sup_y \int f_{Z|X}(y - x|x)\,dx < \infty. \tag{54}$$

Since $f_{Z|X}$ is a pdf for each x, $\sup_x \int f_{Z|X}(y-x|x)\,dy = 1$. For the other, this follows from the assumptions on the variance function:

$$\sup_y \int \frac{1}{\sqrt{2\pi v(x)}} \exp -\frac{(y-x)^2}{2v(x)}\,dx \le \sup_y \int \frac{1}{\sqrt{2\pi v(x)}} \exp -\frac{(y-x)^2}{2v(x)}\,dx \tag{55}$$

$$\le \sup_y \int \frac{1}{\sqrt{2\pi m}} \exp -\frac{(y-x)^2}{2M}\,dx \tag{56}$$

$$= C_{m,M}, \tag{57}$$

where $C_{m,M}$ is a constant depending only on m and M.

Assumption 5. *Satisfied by construction of $\widehat{f}_{X,n}$.*

Assumption 6. *Call $G(x,y) = f_{Z|X}(y-x|x)$. It is sufficient to show that $\mathscr{F} = \{G(\cdot,y)|y \in \mathbb{R}\} \cup \{G(x,\cdot)|x \in \mathbb{R}\}$ is uniformly equicontinuous with modulus $c\varepsilon$, which can be shown by demonstrating that these functions are Lipschitz continuous with constant at most c. For each x, $\partial G/\partial y$ is bounded by $\pm 1/\sqrt{2e\pi v(x)^4}$, so for any x, $G(x,\cdot)$ has Lipschitz constant at most $1/\sqrt{2e\pi m^4}$. Also,*

$$\frac{\partial G(x,y)}{\partial x} = \frac{e^{-\frac{(x-y)^2}{2v(x)}}\left[v'(x)(x-y)^2 - 2v(x)(x-y) - v(x)v'(x)\right]}{\sqrt{8\pi v(x)^5}}, \tag{58}$$

which can be bounded above and below by substituting, depending on the sign of $x-y$, the bounds for $v(x)$ and $v'(x)$. Each function of that form is bounded, with extrema not depending on y, and so each $G(\cdot,y)$ has Lipschitz constant at most the largest magnitude of the extrema of those functions. Thus, the family $\mathscr{F}$ of functions is uniformly equicontinuous with modulus $c\varepsilon$, so $\omega(\varepsilon) = 2c\varepsilon$ is a modulus of continuity satisfying Assumption 6.

Assumption 7. *We note that this requirement is satisfied if $\lambda_n = o(1)$ and $K_n\delta_n^3 \to 0$, the second of which is consistent with the foregoing requirements for satisfying Assumption 2.* □

D.2 Convergence Rate

Assume for simplicity that $m_i = m$ for all n and that the variance function is a constant σ_ϵ^2 and known. Since $m \to \infty$, the outer-level simulations become a negligible portion of the computational cost. Therefore, as in Reference [28], we will use $M = mn$, the number of inner-level simulations, as the computational budget. The best MSE convergence rate for the jackknife estimator in Steckley et al. is $O(M^{-8/11})$ and is achieved with $m = C_1 n^{2/9}$, for some $C_1 > 0$, so $M = C_1 n^{11/9}$. Here, "best rate" means the rate when $m \to \infty$ optimally.

Fan [10] studies kernel deconvolution under the model $Y_i = X_i + \sigma_0\epsilon_i$, where X_i and ϵ_i are mutually independent and each is i.i.d. with distribution independent of n, and $\sigma_0 \to 0$ as $n \to \infty$. Fan's model with the error variance converging to 0 is exactly what is needed to study deconvolution in nested simulation with $m \to \infty$. Fan shows that if $\sigma_0 \to 0$ sufficiently fast, deconvolution estimators can achieve the same convergence rate as when there is no measurement error, i.e., $\sigma_0 = 0$ and a kernel density estimator, not a deconvolution kernel estimator, is used.

To apply Fan's Theorem 4, we make the identifications

$$Y_i = \overline{Y}_i, \quad X_i = X_i, \quad \epsilon_i = \sqrt{m}\,Z_i, \quad \sigma_0 = \frac{1}{\sqrt{m}},$$

where the left-hand side of each equation is the quantity in Fan and the right-hand side is the corresponding quantity in this article. Notice that $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$.

Fan studies convergence in a weighted L_p norm $\|f\|_{wp} = \left\{\int_{-\infty}^{\infty} |f(x)|^p w(x)dx\right\}^{1/p}$, where $1 \le p < \infty$ and w is integrable. If $p = \infty$, then $\|f\|_{wp}$ is the usual sup norm and there is no weighting

function w. The convergence rate that Fan establishes holds for all $p \geq 1$ and, in the case that $p < \infty$, for all integrable w. The rate does not depend on p or w. For comparison with MSE rates in Reference [28], we focus on $\|\cdot\|_{w2}^2$.

The convergence rate of $\hat{f}_X$ depends on the rate at which $\sigma_0 \to 0$ and the order of the kernel. A kernel K has order ν if $\int K(x)x^k dx$ equals 0 for $k = 1, \ldots, \nu - 1$ and is non-zero for $k = \nu$. Assume K is order $\nu = 4$, which requires that K takes both positive and negative values. By Theorem 4 of Fan [10], if $\sigma_0 = C_2 n^{-1/(2\nu+1)} = C_2 n^{-1/9}$, for some $C_2 > 0$, and other assumptions in Fan are met, then the weighted mean squared error of $\hat{f}_X$ converges to 0 at rate $O(n^{-8/9})$. For $\sigma_0 = C_2 n^{-1/9}$ to hold, we need $m = C_1 n^{2/9}$ so $M = C_1 n^{11/9}$ and $n = C_1^{-1} M^{9/11}$ and the weighted squared error rate is $O(M^{-8/11})$, the rate of the jackknife estimator in Steckley et al. [28]. One difference between the results that we derive from Reference [10] and those in Reference [28] is that the latter uses an unweighted MSE, whereas the former cannot use $w \equiv 1$, since this function is not integrable.

A key part of our conjecture is that the QP deconvolution estimator behaves like a kernel deconvolution estimator with a 4th-order kernel. Smoothing a histogram converts a density estimation problem to a regression problem where one smooths the bin areas against the bin centers. Therefore, our estimator is essentially a spline estimator with a piecewise constant spline and a penalty on the second derivative. Li and Ruppert [22] show that spline smoothing with a penalty on the second derivative is asymptotically equivalent to kernel smoothing using a 4th-order kernel. Interestingly, the order of the kernel depends only on the derivative that is penalized, not the degree of the spline.

Our simulation results show that the QP estimator has a smaller MSE than the Steckley jackknife estimator, but not an order of magnitude smaller, suggesting that the difference is in the constants, not the rates, and consistent with our conjecture.

Steckley et al. [28] show that the best rate for the naive estimator is $O(c^{-4/7})$, which, of course, is slower than $O(c^{-8/11})$. The QP deconvolution estimator could achieve a faster convergence rate if the penalty were placed on a derivative higher than the second [34], but whether this would translate into better performance in practical settings is unclear.

ACKNOWLEDGMENTS

The authors gratefully acknowledge detailed feedback from an anonymous Associate Editor and three anonymous Referees, which has helped to improve the article.

REFERENCES

[1] Osei Antwi. 2014. Measuring portfolio loss using approximation methods. *Sci. J. Appl. Math. Statist.* 2, 2 (2014), 42–52. DOI: https://doi.org/10.11648/j.sjams.20140202.11

[2] A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Manag. Sci.* 50, 7 (July 2004), 896–908. Retrieved from https://eprints.soton.ac.uk/48876/.

[3] Russell R. Barton, Henry Lam, and Eunhye Song. 2018. Revisiting direct bootstrap resampling for input model uncertainty. In *Proceedings of the Winter Simulation Conference*. 1635–1645.

[4] Alan Brennan, Samer Kharroubi, Anthony O'Hagan, and Jim Chilcott. 2007. Calculating partial expected value of perfect information via Monte Carlo sampling algorithms. *Med. Decis. Mak.* 27 (2007), 448–470. DOI: https://doi.org/10.1177/0272989X07302555

[5] Mark Broadie, Yiping Du, and Ciamac C. Moallemi. 2011. Efficient risk estimation via nested sequential simulation. *Manag. Sci.* 57 (2011), 1172–1194.

[6] William S. Cleveland and Eric Grosse. 1991. Computational methods for local regression. *Statist. Comput.* 1, 1 (1991), 47–62.

[7] Samuel Daniel Conte and Carl De Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach* (3rd ed.). McGraw-Hill, New York. Retrieved from https://hdl.handle.net/2027/mdp.39015046281518?urlappend=3Bsignon=swle.

[8] Peter J. Diggle and Peter Hall. 1993. A Fourier approach to nonparametric deconvolution of a density estimate. *J. Roy. Statist. Soc. Series B (Methodol.)* (1993), 523–531.

[9] Richard Durrett. 2010. *Probability: Theory and Examples* (4th ed.). Cambridge University Press, Cambridge, UK. Retrieved from http://lib.myilibrary.com/detail.asp?ID=281866.
[10] Jianqing Fan. 1992. Deconvolution with supersmooth distributions. *Canad. J. Statist.* 20 (1992), 155–169.
[11] Simone Farinelli and Mykhaylo Shkolnikov. 2012. Two models of stochastic loss given default. *J. Cred. Risk* 8, 2 (June 2012), 3–20. DOI: https://doi.org/10.21314/JCR.2012.141
[12] M. Gordy and S. Juneja. 2006. Efficient simulation for risk measurement in portfolio of CDOs. In *Proceedings of the Winter Simulation Conference*. 749–756. DOI: https://doi.org/10.1109/WSC.2006.323155
[13] Michael B. Gordy and Sandeep Juneja. 2010. Nested simulation in portfolio risk measurement. *Manag. Sci.* 56, 10 (2010), 1833–1848. DOI: https://doi.org/10.1287/mnsc.1100.1213
[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The elements of statistical learning: prediction, inference and data mining. Springer-Verlag, New York.
[15] Rouba Ibrahim and Pierre L'Ecuyer. 2013. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manuf. Serv. Oper. Manag.* 15 (02 2013), 72–85. DOI: https://doi.org/10.1287/msom.1120.0405
[16] Rouba Ibrahim, Han Ye, Pierre L'Ecuyer, and Haipeng Shen. 2016. Modeling and forecasting call center arrivals: A literature survey and a case study. *Int. J. Forecast.* 32, 3 (2016), 865–874. DOI: https://doi.org/10.1016/j.ijforecast.2015.11.012
[17] Gabriel Jiménez and Javier Mencía. 2009. Modelling the distribution of credit losses with observable and latent factors. *J. Empir. Finan.* 16, 2 (2009), 235–253. DOI: https://doi.org/10.1016/j.jempfin.2008.10.003
[18] Hai Lan, Barry L. Nelson, and Jeremy Staum. 2010. A confidence interval procedure for expected shortfall risk measurement via two-level simulation. *Oper. Res.* 58, 5 (2010), 1481–1490. DOI: https://doi.org/10.1287/opre.1090.0792
[19] Peter D. Lax. 2002. *Functional Analysis*. Wiley, New York. Retrieved from http://newcatalog.library.cornell.edu/catalog/4301419.
[20] Shing-Hoi Lee. 1998. *Monte Carlo Computation of Conditional Expectation Quantiles*. Ph.D. Dissertation. Stanford University, Stanford, CA.
[21] Shing-Hoi Lee and Peter W. Glynn. 2003. Computing the distribution function of a conditional expectation via Monte Carlo: Discrete conditioning spaces. *ACM Trans. Model. Comput. Simul.* 13, 3 (July 2003), 238–258. DOI: https://doi.org/10.1145/937332.937334
[22] Yingxing Li and David Ruppert. 2008. On the asymptotics of penalized splines. *Biometrika* 95, 2 (2008), 415–436.
[23] Pierre L'Ecuyer and Eric Buist. 2006. Variance Reduction in the Simulation of Call Centers. 604–613. DOI: https://doi.org/10.1109/WSC.2006.323136
[24] John Mendelsohn and John Rice. 1982. Deconvolution of microfluorometric histograms with B splines. *J. Amer. Statist. Assoc.* 77, 380 (1982), 748–753. DOI: https://doi.org/10.2307/2287301
[25] Boris Oreshkin, Nazim Regnard, and L'Ecuyer. 2016. Rate-based daily arrival process models with application to call centers. *Oper. Res.* 64 (03 2016). DOI: https://doi.org/10.1287/opre.2016.1484
[26] Murray H. Protter and Charles B. Morrey, Jr. 1985. *Intermediate Calculus (2nd ed.)*. Springer, New York, 421–426.
[27] R Core Team. 2018. R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/.
[28] Samuel G. Steckley, Shane G. Henderson, David Ruppert, Ran Yang, Daniel W. Apley, and Jeremy Staum. 2016. Estimating the density of a conditional expectation. *Electron. J. Statist.* 10, 1 (2016), 736–760. DOI: https://doi.org/10.1214/16-EJS1121
[29] L. A. Stefanski and R. J. Carroll. 1990. Deconvoluting kernel density estimators. *Statistics* 21 (1990), 169–184.
[30] Yunpeng Sun, Daniel W. Apley, and Jeremy Staum. 2011. Efficient nested simulation for estimating the variance of a conditional expectation. *Oper. Res.* 59, 4 (2011), 998–1007. Retrieved from http://www.jstor.org/stable/23013161.
[31] János Sztrik. 2011. *Basic Queueing Theory*. GlobeEdit.
[32] L. Takács. 1962. *Introduction to the Theory of Queues*. Oxford University Press.
[33] A. W. van der Vaart. 1998. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK. Retrieved from http://dx.doi.org/10.1017/CBO9780511802256
[34] Xiao Wang, Jinglai Shen, and David Ruppert. 2009. Local asymptotics of P-spline smoothing. *arXiv preprint arXiv:0912.1824* (2009).
[35] Wei Xie, Barry L. Nelson, and Russell R. Barton. 2014. A Bayesian framework for quantifying uncertainty in stochastic simulation. *Oper. Res.* 62, 6 (2014), 1439–1452. DOI: https://doi.org/10.1287/opre.2014.1316
[36] Ran Yang, Daniel Apley, Jeremy Staum, and David Ruppert. 2020. Density deconvolution with additive measurement errors using quadratic programming. *J. Computat. Graphic. Statist.* (2020).
[37] L. C. Zhao, P. R. Krishnaiah, and X. R. Chen. 1991. Almost Sure L_r-Norm Convergence for Data-Based Histogram Density Estimates. *Theor. Probab. Applic.* 35, 2 (Jan. 1991), 396–403. DOI: https://doi.org/10.1137/1135057

Received November 2019; revised September 2020; accepted March 2021