

DIMENSION REDUCTION AND ORACLE OPTIMALITY IN CONDITIONAL DENSITY ESTIMATION AND ACTUARIAL APPLICATION

Sam Efromovich ¹

In many applications, ranging from civil engineering and economics to life and actuarial sciences, knowing the conditional density of response given predictors (covariates, explanatory variables) allows one to solve a wide variety of problems including prediction, hypothesis testing, learning and discriminant analysis. In applications it is typical to have mixed (continuous and ordinal/nominal categorical) covariates and a relatively small sample size. In general, this presents a major obstacle due to familiar curse of dimensionality. At the same time, it is often the case that response is conditionally independent of some of the covariates. Then the statistician must seize the opportunity of dimension reduction to attenuate the curse. Further, it is always prudent to employ an optimal estimation. This article suggests a nonparametric estimator that performs the wished dimension reduction and optimal estimation via mimicking an oracle that knows an underlying conditional density. The main theoretical result is an oracle inequality which relates oracle's and estimator's mean integrated squared errors for any underlying conditional density. A numerical study, motivated by Simpson's paradox, sheds light on performance of the estimator for small samples. Then the estimator is used to analyze the actuarial practice of using credit score as a rating variable in calculation of auto-insurance premiums.

KEY WORDS: Asymptotic; Continuous and categorical data; Curse of multidimensionality; Credit Score; Nonparametric; Oracle inequality; Simpson's paradox; Small sample.

¹Sam Efromovich is Professor, Department of Mathematical Sciences, Univ. of Texas at Dallas, Richardson, TX, 75083-0688 (E-mail: *efrom@utdallas.edu*). This work was partially supported by NSF grant DMS-0604558 and actuarial grant TAF/CAS-07.

1. INTRODUCTION

Consider the case of a controlled experiment with objective to understand a relationship between an explanatory data vector X (so-called predictor) and a univariate variable of interest (so-called response) Y . In general the predictor may have mixed covariates. To be specific, similarly to the JASA article HRL (Hall, Racine and Li 2004), which will be used as the main reference/benchmark, it is assumed that $X = (U_1, \dots, U_m, V)$ has m continuous covariates and a single categorical covariate V which takes on a finite number of values coded as $0, 1, \dots, m'$ (note that by an appropriate coding any vector of categorical covariates can be collapsed into a single covariate). Observations are i.i.d. samples $(X_1, Y_1), \dots, (X_n, Y_n)$ from (X, Y) with a joint probability density $p(x)f(y|x)$ where predictors are distributed according to a design density $p(x) = p(u_1, \dots, u_m|v)p^*(v)$ with $p^*(v)$ being the probability mass function of the categorical covariate; see a discussion in HRL. The problem is to estimate the underlying conditional density $f(y|x)$ without any assumption about its shape and/or smoothness, that is using a nonparametric approach. Remember that while it is tempting to convert the problem into estimation of the joint multivariate density, this does not lead to an optimal solution as lower bounds in Efromovich (2007) indicate. In other words, estimation of conditional and joint densities, while having a lot in common, requires special methods and methodologies.

It is well known that the problem of nonparametric conditional density estimation is complicated by the curse of dimensionality. HRL, on page 1015, presented the following nice explanation of the curse which motivated their research. Assume that the conditional density is twice differentiable in each continuous variable. Then a conventional (and popular in the literature) second-order kernel estimator converges at rate $n^{-2/(m+5)}$ which is dramatically slower than the rate $n^{-2/5}$ of estimation of the univariate marginal density of the response. Further, in practical applications the curse necessitates to use dramatically larger sample sizes.

In general, there is no way to overcome this curse apart of using larger sample sizes. At the same time, if the conditional density $f(y|u_1, \dots, u_m, v)$ depends only on $m_1 < m$ continuous covariates (with Y being conditionally independent from the other $m - m_1$ continuous covariates) then the rate improves to $n^{-2/(m_1+5)}$.

The HRL's attenuation of "irrelevant" (using the HRL's terminology) covariates belongs to a

broader (and currently hot) topic of dimension reduction. Dimension reduction is looking after opportunities which may or may not exist. Further, there exists a classical statistical approach for dealing with the curse by employing an optimal multidimensional estimation procedure. Let us comment on this approach using the above-formulated example of HRL. Suppose that indeed only m_1 continuous covariates are “relevant” and that a perfect adaptive second-order kernel estimator (for instance the one developed in HRL) attains the wished rate $n^{-2/(m_1+5)}$. Note that this is not only the wished but also the fastest rate attained by the kernel estimator regardless of how smooth the conditional density is. Further, the rate $n^{-2/(m_1+5)}$ is optimal only if the density is at most twice differentiable in each continuous variable. There are many practical examples where the conditional density is smoother, for instance in a classical polynomial regression with additive normal errors where the conditional density is infinitely-fold differentiable with respect to the response. To shed light on possible rates in conditional density estimation, let us make a simple calculation. If the conditional density is at least $2m + 3$ times differentiable in each continuous variable, then an estimator can converge with the minimax rate $n^{-(2m+3)/(5m+7)}$ which, for any pair (m_1, m) , is faster than the benchmark rate $n^{-2/(m_1+5)}$ of the HRL’s kernel estimator.

We may conclude that in the analysis of conditional densities it is highly desirable to combine the idea of optimal estimation with the idea of dimension reduction. Developing of the corresponding methodology and methods will be the main goal of this article. Further, a complementary aim of the paper is to relax the HRL’s assumption about conventional independence between a group of covariates and the response, and replace it by conditional independence.

To solve the formulated problems, it is suggested to develop the methodology of mimicking an oracle which performs an ideal dimension reduction in the sense of HRL and, at the same time, delivers adaptive, to any underlying smoothness, estimation of the conditional density. Oracle is a pseudo-estimator that knows both data and the conditional density, and the art of choosing an appropriate oracle is in finding a compromise between its statistical properties and the possibility of its mimicking by a data-driven estimator. An interesting discussion of this and other approaches can be found in Simonoff (1996), Hart (1997), Johnstone (1998), Eubank (1999), Yang (2000), Samarov and Tsybakov (2005), Wasserman (2005) and Efromovich (2007).

The content of this article is as follows. Section 2 introduces the oracle methodology of es-

timization of the conditional density with mixed covariates. Section 3 explains how to develop a data-driven estimator which mimics the recommended oracle, and Section 4 presents the main theoretical result — an oracle inequality — which establishes how well the data-driven nonparametric estimator matches its benchmark. Numerical study and an actuarial example can be found in Sections 5 and 6. Proofs are given in the Appendix.

2. ORACLE METHODOLOGY

Without any loss of generality, let us restrict our attention to the case $m = 2$ with $x := (u_1, u_2, v)$. Remember that the regression model was introduced in the Introduction. Following HRL, it is assumed that $z := (y, u_1, u_2, v) \in [0, 1]^3 \times \{0, 1, \dots, m'\}$. Introduce: an orthonormal basis $\{\varphi_j(y), j = 0, 1, \dots\}$ on $[0, 1]$ for the response variable y ; an orthonormal basis $\{\psi_r(u), r = 0, 1, \dots\}$ on $[0, 1]$ for a continuous covariate; an orthonormal basis $\{\phi_t(v), t = 0, 1, \dots, m'\}$ with equal weights $(m' + 1)^{-1}$ on $\{0, 1, \dots, m'\}$ for the categorical variable v . For now the only property of elements of the bases that we need to suppose is that they are bounded and $\varphi_0(y) = \psi_0(u) = \phi_0(v) = 1$ for all y, u and v . There are many bases to choose from; see examples in Hall (1983) and Efromovich (1999).

Let for each value of v the conditional density $f(y|u_1, u_2, v)$, as a function in the triplet (y, u_1, u_2) , be square integrable on $[0, 1]^3$. Then we can define Fourier coefficients (corresponding to the tensor-product basis)

$$\theta_{jrst} := (m' + 1)^{-1} \sum_{v=0}^{m'} \int_{[0,1]^3} f(y|u_1, u_2, v) \varphi_j(y) \psi_r(u_1) \psi_s(u_2) \phi_t(v) dy du_1 du_2, \quad (1)$$

which allow us to write the conditional density as the following Fourier expansion

$$f(y|x, u_1, u_2, v) = \sum_{t=0}^{m'} \sum_{j,r,s=0}^{\infty} \theta_{jrst} \varphi_j(y) \psi_r(u_1) \psi_s(u_2) \phi_t(v). \quad (2)$$

Now let us consider an oracle that knows how to take advantage of the following hierarchy of dimension-reduction possibilities. The most attractive possibility is where the response Y is independent of all covariates, that is $f(y|u_1, u_2, v) \equiv f(y)$ for all values of the covariates, and for an oracle a multidimensional estimation problem is replaced by a univariate one. In this case, due to our choice of the tensor-product basis,

$$f(y|u_1, u_2, v) = \sum_{j=0}^{\infty} \theta_{j000} \varphi_j(y) = \sum_{j=0}^{\infty} \theta_j \varphi_j(y), \quad (3)$$

where $\theta_j := \int_0^1 f(y)\varphi_j(y)dy$ is the Fourier coefficient of the marginal density of the response. Given such an opportunity, an adaptive (to the underlying smoothness of f) univariate oracle, for instance the blockwise-shrinkage oracle of Efromovich (1985), can be utilized and then successfully mimicked by a data-driven univariate estimator; see a discussion of this approach in Efromovich (1999). Of course, the use of a univariate estimator should be hinted by an oracle, and this is the major statistical challenge.

The above-considered case is the best statistical opportunity. Next one in our hierarchy of the statistical “luck” is when the response, given a particular covariate, conditionally independent of other covariates. As an example, let this covariate be U_1 and then $f(y|u_1, u_2, v) \equiv f(y|u_1)$ for all possible values of (u_2, v) . In this case an oracle will estimate a bivariate function

$$f(y|u_1, u_2, v) = f(y|u_1) = \sum_{j,r=0}^{\infty} \theta_{jr00}\varphi_j(y)\psi_r(u_1) = \sum_{j,r=0}^{\infty} \theta_{jr}\varphi_j(y)\psi_r(u_1), \quad (4)$$

where $\theta_{jr} = \int_{[0,1]^2} f(y|u_1)\varphi_j(y)\psi_r(u_1)dydu_1$ are Fourier coefficients of $f(y|u_1)$.

By induction, it is reasonable to suggest that an oracle could use the following expansion:

$$\begin{aligned} f(y|u_1, u_2, v) &= \left[\sum_{j=0}^{\infty} \theta_{j000}\varphi_j(y) \right] \\ &+ \left[\sum_{j=1}^{\infty} \sum_{t=1}^{m'} \theta_{j00t}\varphi_j(y)\phi_t(v) + \sum_{j,r=1}^{\infty} \theta_{jr00}\varphi_j(y)\psi_r(u_1) + \sum_{j,s=1}^{\infty} \theta_{j0s0}\varphi_j(y)\psi_s(u_2) \right] \\ &+ \left[\sum_{j,r=1}^{\infty} \sum_{t=1}^{m'} \theta_{jr0t}\varphi_j(y)\psi_r(u_1)\phi_t(v) + \sum_{j,s=1}^{\infty} \sum_{t=1}^{m'} \theta_{j0st}\varphi_j(y)\psi_s(u_2)\phi_t(v) \right. \\ &\left. + \sum_{j,r,s=1}^{\infty} \theta_{jrs0}\varphi_j(y)\psi_r(u_1)\psi_s(u_2) \right] + \left[\sum_{j,r,s=1}^{\infty} \sum_{t=1}^{m'} \theta_{jrst}\varphi_j(y)\psi_r(u_1)\psi_s(u_2)\phi_t(v) \right]. \quad (5) \end{aligned}$$

The reader might note that some terms of expansion (2) are skipped in (5) because $\int_0^1 f(y|u_1, u_2, v)dy \equiv 1$ implies $\theta_{0rst} = 0$ for all $r + s + t > 0$. This simplification will be very handy in constructing estimator for small sample sizes.

Oracle’s expansion (5) exhibits 4 distinct groups of Fourier terms highlighted by square brackets. The first one is identical to (3) and presents the most beneficial for dimension-reduction case where the response is independent of the multivariate predictor. The second group is motivated by (4), and here each sum corresponds to the case where 2 covariates are conditionally independent of the

response given another covariate. Please note that the first sum in the second group presents a more favorable dimension-reduction opportunity where, given the categorical covariate, the response is conditionally independent of the both continuous covariates. If the latter is indeed the case, that is $f(y|u_1, u_2, v) = f(y|v)$, then an oracle can achieve a univariate rate of estimation. The third group of Fourier terms corresponds to the less favorable dimension-reduction case where an oracle explores a possibility that the response depends only on two covariates. Similarly to the previous case, among three sums in this group, the first two imply the possibility of more accurate estimation whenever an estimated conditional density depends only on one continuous covariate. Finally, terms of the last group appear when no dimension reduction is possible.

Now we are in a position to define an oracle that performs the above-described dimension reduction and that can be mimicked by a data-driven estimator. The oracle makes the following steps. First of all, in (5) the underlying Fourier coefficients θ_{jrst} are replaced by their sample mean estimates

$$\tilde{\theta}_{jrst} := n^{-1}(m' + 1)^{-1} \sum_{l=1}^n \varphi_j(Y_l) \psi_r(U_{1l}) \psi_s(U_{2l}) \phi_t(V_l) p^{-1}(U_{1l}, U_{2l}, V_l). \quad (6)$$

Recall that $p(u_1, u_2, v)$ is a known design density. Second, a blockwise shrinkage of Fourier coefficients within each of 8 sums in (5) is utilized. Namely, let B_{ik} denote a k th block of indexes in i th sum. Then a blockwise shrinkage oracle (which is motivated by a famous Wiener filter) multiplies all empirical Fourier coefficients (6) from the block by a factor

$$\mu_{ik} := \frac{\Theta_{ik}}{L_{ik}^{-1} E\{\sum_{(j,r,s,t) \in B_{ik}} \tilde{\theta}_{jrst}^2\}}. \quad (7)$$

Here

$$\Theta_{ik} := L_{ik}^{-1} \sum_{(j,r,s,t) \in B_{ik}} \theta_{jrst}^2, \quad (8)$$

and L_{ik} denotes the cardinality of B_{ik} (number of indexes belonging to the block). Note that for our particular expansion (5) we have 8 arrays of blocks $\{B_{ik}, i = 1, 2, \dots, 8, k = 1, 2, \dots\}$, and we also introduce 8 cutoffs $K_i, i = 1, 2, \dots, 8$. In general, elements of the arrays may depend on the sample size n , and cutoffs always (implicitly) depend on n , but none depends on observations.

Then a blockwise-shrinkage oracle is defined as

$$\tilde{f}^*(y|u_1, u_2, v) = \sum_{i=1}^8 \sum_{k=1}^{K_i} \mu_{ik} \sum_{(j,r,s,t) \in B_{ik}} \tilde{\theta}_{jrst} \varphi_j(y) \psi_r(u_1) \psi_s(u_2) \phi_t(v). \quad (9)$$

Note that $\tilde{\theta}_{jrst}$ is an unbiased estimate of θ_{jrst} . As a result $E\{\tilde{\theta}_{jrst}^2\} \geq \theta_{jrst}^2$ and this yields $0 \leq \mu_{ik} \leq 1$. The latter explains why the oracle is called blockwise-shrinkage.

Further, it is easy to see from (7)-(9) that if i th sum in (5) is zero then $\mu_{ik} = 0$ for all $k = 1, 2, \dots$ and the i th sum in (9) is also zero. This yields the wished dimension reduction property of the oracle. Further, as we shall see shortly, the oracle has excellent asymptotic properties over standard function classes discussed in the literature. The only remaining question is the possibility to suggest an estimator that is able to mimic this oracle; this issue is discussed in next sections.

Remark 1. If the categorical covariate is a combination of several underlying categorical covariates, then similarly to the case of continuous covariates one can consider them as a vector-covariate. This will be benefit the estimator if some categorical covariates are conditionally independent of the response; however this approach cannot improve rates.

3. ESTIMATOR

The suggested estimator mimics Wiener's shrinkage (7) by a plug-in procedure

$$\tilde{\mu}_{ik} := \frac{\hat{\Theta}_{ik}}{\tilde{\Theta}_{ik}} I(\hat{\Theta}_{ik} > t_{ik} n^{-1}), \quad (10)$$

where t_{ik} is a positive threshold level,

$$\begin{aligned} \hat{\Theta}_{ik} := & L_{ik}^{-1} [2/(n(n-1))] (m' + 1)^{-2} \sum_{1 \leq l < q \leq n} \sum_{(j,r,s,t) \in B_{ik}} \varphi_j(Y_l) \psi_r(U_{1l}) \psi_s(U_{2l}) \phi_t(V_l) p^{-1}(U_{1l}, U_{2l}, V_l) \\ & \times \varphi_j(Y_q) \psi_r(U_{1q}) \psi_s(U_{2q}) \phi_t(V_q) p^{-1}(U_{1q}, U_{2q}, V_q) \end{aligned} \quad (11)$$

is an unbiased estimate of the Sobolev functional Θ_{ik} defined in (8), and

$$\tilde{\Theta}_{ik} := L_{ik}^{-1} \sum_{(j,r,s,t) \in B_{ik}} \tilde{\theta}_{jrst}^2 \quad (12)$$

is an unbiased estimate of the denominator in (7).

Then, following (9), an estimator of the conditional density is defined,

$$\tilde{f}(y|u_1, u_2, v) = \sum_{i=1}^8 \sum_{k=1}^{K_i} \tilde{\mu}_{ik} \sum_{(j,r,s,t) \in B_{ik}} \tilde{\theta}_{jrst} \varphi_j(y) \psi_r(u_1) \psi_s(u_2) \phi_t(v). \quad (13)$$

Next section presents an oracle inequality which shows how well this estimator matches the oracle.

Remark 2. In the case of an unknown design density (passive experiment), the design density should be estimated and then plugged in. This is a known approach discussed, for instance, in Efromovich (1999,2007). Further, Section 5 presents an example of a passive experiment with extra observations of covariates that can be used to estimate the design density. Such a situation often occurs in studies where responses are more difficult/costly to obtain.

4. ORACLE INEQUALITY

To make formulae simpler, introduce specific bases for continuous covariates. A classical trigonometric basis $\{\varphi_0(y) = 1, \varphi_{2j-1}(y) = 2^{1/2} \sin(2\pi jy), \varphi_{2j}(y) = 2^{1/2} \cos(2\pi jy), j = 1, 2, \dots\}$ is used for the response and a classical cosine basis $\{\psi_0(u) = 1, \psi_r(u) = 2^{1/2} \cos(\pi ru), r = 1, 2, \dots\}$ is used for a continuous covariate. In what follows $n > 3$.

Assumption 1. *It is assumed that $f(y|u_1, u_2, v) < C_1 < \infty$ and the conditional density is integrable on $[0, 1]^3$, and $p(u_1, u_2, v) > C_2 > 0$.*

Depending on a particular proposition, it may be convenient to introduce some restrictions on used cutoffs, thresholds and blocks. First of all, let us recall that even for the case of estimation of a univariate differentiable probability density the order of the number of estimated Fourier coefficients is at most $n^{-1/3}$. This allows us to introduce the following restriction,

$$\max_{i \in \{1, \dots, 8\}} \max_{k \leq K_i} L_{ik} n^{-1/3} < C^* < \infty. \quad (14)$$

Further, to simplify formulae it is assumed that blocks are such that for all $l = 1, 2, \dots$ and any (r, s, t, i, k)

$$(2l, r, s, t) \in B_{ik} \Leftrightarrow (2l - 1, r, s, t) \in B_{ik}. \quad (15)$$

Let us introduce the main assumption on blocks and thresholds which is similar to the minimal one known for the case of univariate functions; see a discussion in Efromovich (2004).

Assumption 2. *Let $\max_{i \in \{1, \dots, 8\}} t_{ik} \rightarrow 0$ as $k \rightarrow \infty$ and for any positive constant a*

$$\sum_{i=1}^8 \sum_{k=1}^{K_i} L_{ik}^{3/4} \exp(-at_{ik}^2 L_{ik}) < \infty. \quad (16)$$

Theorem 1. (Oracle Inequality.) *Suppose that Assumptions 1-2 and (14)-(15) hold. Then for any $c \in (0, 1)$ the following inequality relates risks of the estimator \tilde{f} defined in (13) and the*

oracle \tilde{f}^* defined in (9),

$$\begin{aligned}
& (m' + 1)^{-1} E \left\{ \sum_{v=0}^{m'} \int_{[0,1]^3} (\tilde{f}(y|u_1, u_2, v) - f(y|u_1, u_2, v))^2 dy du_1 du_2 \right\} \\
\leq & (1+c)(1+\delta_n(f))(m'+1)^{-1} E \left\{ \sum_{v=0}^{m'} \int_{[0,1]^3} (\tilde{f}^*(y|u_1, u_2, v) - f(y|x_1, x_2, v))^2 dy du_1 du_2 \right\} + (1+c^{-1})C_* n^{-1},
\end{aligned} \tag{17}$$

where $C_* < \infty$ and $\delta_n(f) < C < \infty$. Further, if an estimated conditional density is nonparametric in the sense that the oracle's risk vanishes slower than n^{-1} , that is

$$n E \left\{ \sum_{v=0}^{m'} \int_{[0,1]^3} (\tilde{f}^*(y|u_1, u_2, v) - f(y|u_1, u_2, v))^2 dy du_1 du_2 \right\} \rightarrow \infty, \quad n \rightarrow \infty,$$

then $\delta_n(f) \rightarrow 0$ as $n \rightarrow \infty$.

Remember that n^{-1} is the parametric rate of convergence. Thus the remainder term in the oracle inequality cannot be smaller in order. This, together with the minimal Assumption 2 on blocks and thresholds, highlights the sharpness of the oracle inequality.

Further, remember that the oracle performs the wished dimension reduction whenever such an opportunity occurs. This together with Theorem 1 yield the following important proposition.

Corollary 1. *Suppose that the assumption of Theorem 1 holds. Then risks of the estimator and the oracle decrease with the same rate as the sample size increases. Further, if an estimated conditional density is nonparametric, then the ratio of the risks tends to 1 as the sample size increases. As a result, the estimator simultaneously performs the wished dimension reduction and matching the oracle's risk.*

Now let us calculate the oracle's risk.

Theorem 2. (Oracle's risk.) *Suppose that Assumption 1 and (15) hold. Then for any arrays of blocks, thresholds and cutoffs the oracle's risk is calculated by the formula*

$$\begin{aligned}
& (m' + 1)^{-1} E \left\{ \sum_{v=0}^{m'} \int_{[0,1]^3} (\tilde{f}^*(y|u_1, u_2, v) - f(y|u_1, u_2, v))^2 dy du_1 du_2 \right\} \\
& = \sum_{i=1}^8 n^{-1} \sum_{k=1}^{K_i} L_{ik} \mu_{ik} (D_{ik} - \Theta_{ik}) + \sum_{i=1}^8 \sum_{k > K_i} L_{ik} \Theta_{ik},
\end{aligned} \tag{18}$$

where μ_{ik} and Θ_{ik} are defined in (7)-(8) and

$$D_{ik} := L_{ik}^{-1}(m' + 1)^{-2} \sum_{(j,r,s,t) \in B_{ik}} \sum_{v=0}^{m'} \int_{[0,1]^2} \psi_r^2(u_1) \psi_s^2(u_2) \phi_t^2(v) p^{-1}(u_1, u_2, v) du_1 du_2. \quad (19)$$

Further,

$$\mu_{ik} = \frac{\Theta_{ik}}{\Theta_{ik} + n^{-1}(D_{ik} - \Theta_{ik})}, \quad (20)$$

$$D_{ik} - \Theta_{ik} > 0, \quad (21)$$

and if the design of covariates is uniform (meaning that $p(u_1, u_2, v) \equiv (m' + 1)^{-1}$) then

$$D_{ik} \equiv 1. \quad (22)$$

Let us note that so far all results have been obtained under a pointwise approach for a specific underlying conditional density f . At the same time, formula (18), together with a standard approach of Efromovich (1999,2000,2007), implies that the oracle is sharp minimax over a vast set of Sobolev, analytic and entire function classes, as well as that it is superefficient in the sense of Brown *et al.* (1997). Then Corollary 1 yields that the estimator has the same statistical properties. There is no other conditional density estimator, suggested in the literature, that has such pointwise and global statistical properties; see a discussion in HRL, Hydman, Bashtannyk and Grunwald (1996), Hydman and Yao (2002), and Hall and Yao (2005).

Let us finish this section by presenting a technical result which is instrumental in establishing the oracle inequality and which is of interest on its own because it describes basic statistical properties of the Sobolev statistic $\hat{\Theta}$ (an unbiased estimate of the Sobolev functional Θ defined in (11)).

Lemma 1. *Let Assumption 1 hold. Consider a particular block B_{ik} whose length satisfies $L_{ik}n^{-1} \leq C^* < \infty$. Then the following moment inequality holds:*

$$E\{(\hat{\Theta}_{ik} - \Theta_{ik})^2\} \leq C_3 L_{ik}^{-1} n^{-1} (L_{ik}^{1/2} \Theta_{ik} + n^{-1}), \quad (23)$$

where the finite constant C_3 depends only on C_1 , C_2 and C^* . Further, for any positive constant κ the following exponential inequality holds:

$$\Pr(|\hat{\Theta}_{ik} - \Theta_{ik}| > \kappa t_{ik} n^{-1})$$

$$\begin{aligned} &\leq C_4 \left[\exp\{-C_4^{-1} \min[\kappa^2 t_{ik}^2 L_{ik}, \kappa t_{ik} L_{ik}, n^{1/3} \kappa^{2/3} t_{ik}^{2/3} L_{ik}^{1/6}, (\kappa t_{ik} n)^{1/2}]\} \right. \\ &\quad \left. + \exp\{-C_4^{-1} \frac{\kappa^2 t_{ik}^2 L_{ik}^{1/2}}{n \Theta_{ik} + \kappa t_{ik} (L_{ik} \Theta_{ik})^{1/2}}\} + \exp\{-C_4^{-1} \frac{\kappa^2 t_{ik}^2 L_{ik} n}{1 + \kappa t_{ik} L_{ik}}\} \right], \end{aligned} \quad (24)$$

where the finite positive constant C_4 depends only on C_1 , C_2 and C^* .

5. NUMERICAL STUDY

It is well known that any nonparametric estimation requires relatively large sample sizes; see a discussion and examples in Efromovich (1999). Let us explore a possible quality of conditional density estimation via an extension of Simpson's paradox known for a two-way table setting. Here we are using the NSF example discussed in section 9.1 of Moore and McCabe (2003).

A study, based on the analysis of salary (the response variable Y) of 100 male and 100 female engineers and scientists within 10 years since receiving the Master's degree, found that the mean salary of female respondents was only 78% of the mean salary for male respondents. Further, if the time since graduation (the covariate X) is taken into account, then regression lines, shown in Figure 1, indicate that the gap between those salaries increases in time. Note that all observations are rescaled onto unit square. A two-way table Simpson's paradox is that if we take into account the field (engineering or science), then the mean salaries are 0.55, 0.58, 0.34 and 0.39 for male-engineers, female-engineers, male-scientists and female-scientists, respectively. In other words, women do better than men in every field, and yet fall behind men when we aggregate the data.

Figure 2 allows us to appreciate a regression analog of the Simpson's paradox. Regressions indicate that, in both fields and at any time since graduation, women's salaries are higher. Further, women's salary increases faster in time. As we see, the regression analog of Simpson's paradox is even more confusing than the traditional one known for two-way table where just mean values are compared.

The observations were generated by a model $Y = 2 + I(V \leq 1) + [2I(V \leq 1) + I(V = 1) + I(V = 3)]X + 0.5\epsilon$. Here ϵ is a standard normal regression error, $V = 0$ for male-engineers, $V = 1$ for female-engineers, $V = 2$ for male-scientists, and $V = 3$ for female-scientists, $P(V = 0) = P(V = 3) = 0.4$ and $P(V = 1) = P(V = 2) = .1$, and X is uniformly distributed on $[0, 1]$. After all Y s are recorded, they are rescaled onto $[0, 1]$.

Now let us check if a conditional density estimator can help one to realize the source of the paradox. The total sample size $n = 200$ is obviously small, and this fact should be taken into account. Because the estimator of Section 3 is a blockwise estimator mimicking corresponding Efromovich-Pinsker oracle, construction of its small-data counterpart can follow the general methodology of Efromovich (1999, s.3.3,6.6). Here and in what follows the estimator uses cosine bases and default parameters $cJ0 = 4, cJ1 = .5, cJM = 1, cT = 2, cB = 2$ explained in the book.

Two top diagrams in Figure 3 exhibit the estimated conditional densities of salary given time since graduation and gender (compare with underlying densities shown below). Each estimate indicates two pronounced ridges that may be explained by a lurking variable. As a result, visualization of the conditional density may help one in understanding the data, making a correct assumption about an underlying model, and thus be “prepared” for Simpson’s paradox. We can conclude that visualization of the conditional density sheds a new light on data and can be a handy statistical tool in addition to classical regression.

Figure 4 shows us estimated conditional densities when field is taken into consideration. Here the nominal categorical variable takes on 4 values depending on gender and field. The estimates are far from perfect but reasonable given sample sizes exhibited in Figure 2.

Now let us use the above-described model of Simpson’s paradox to conduct an intensive Monte Carlo study of the conditional density estimator. In addition to the case of one continuous covariate X , we will also consider two settings with an extra covariate Z , say the (rescaled) age at graduation. In the first setting $Z = Z1$ with $Z1$ being independent of (Y, X) and uniformly distributed. Remember that this is the type of “irrelevance” of an extra covariate studied in HRL. In the second setting $Z = Z2$ with $Z2$ being exponentially distributed with rate $1/(.3 + X)$ (that is, its mean is $.3 + X$) and then truncated on $[0, 1]$. Note that X and $Z2$ are collinear, Y and $Z2$ are dependent but they are conditionally independent given X . According to the theory of Section 4, the estimator should recognize both the independence and the conditional independence, and then mimic an oracle that knows these facts. Let us check if the estimator can do this for small sample sizes.

Table 1. Results of a numerical study for the Simpson's paradox

		$f(y x, v)$	$f(y x, z1, v)$		$f(y x, z2, v)$	
n	V	ISE/ISEO	ISE/ISEO	% indep	ISE/ISEO	% indep
100	0	1.18	1.35	41	1.42	33
	1	1.16	1.33	44	1.41	36
	2	1.19	1.28	44	1.45	36
	3	1.21	1.33	41	1.42	34
	4	1.17	1.27	45	1.34	41
	5	1.15	1.23	45	1.32	41
200	0	1.13	1.27	52	1.38	44
	1	1.12	1.25	57	1.33	52
	2	1.12	1.24	56	1.34	54
	3	1.15	1.27	51	1.38	44
	4	1.12	1.23	60	1.29	56
	5	1.15	1.21	59	1.28	56
400	0	1.08	1.13	73	1.16	64
	1	1.11	1.14	64	1.19	60
	2	1.11	1.18	64	1.26	60
	3	1.06	1.21	61	1.25	58
	4	1.07	1.11	71	1.19	65
	5	1.08	1.03	71	1.19	65
600	0	1.02	1.04	65	1.06	57
	1	1.04	1.08	61	1.09	56
	2	1.13	1.14	58	1.23	53
	3	1.07	1.10	60	1.11	58
	4	1.01	1.03	77	1.06	64
	5	1.03	1.06	77	1.10	63
800	0	1.02	1.04	67	1.08	60
	1	1.04	1.07	63	1.08	60
	2	1.11	1.15	63	1.17	57
	3	1.05	1.06	67	1.06	61
	4	1.02	1.06	79	1.09	73
	5	1.01	1.04	79	1.08	73
1000	0	1.01	1.03	77	1.05	75
	1	1.06	1.05	79	1.06	75
	2	1.10	1.12	79	1.14	74
	3	1.04	1.06	77	1.09	75
	4	1.01	1.02	80	1.05	76
	5	1.01	1.03	80	1.06	76

In our numerical study we evaluate 3 densities: $f(y|x, v)$, $f(y|x, z1, v)$ and $f(y|x, z2, v)$. Cate-

gorical variable V takes either 4 values $\{0, 1, 2, 3\}$ corresponding to gender-field or just two values corresponding to gender. To simplify references, in the latter case of V being gender, we denote its values as 4 and 5 for male and female graduate, respectively. For each experiment (which includes a specific sample size n and one of the 3 studied conditional densities), a thousand of simulations is conducted and then integrated squared errors (ISEs) are calculated.

Results of the study are presented in Table 1. The first two columns present the sample size and the considered value of V (remember that values 0-3 and 4-5 describe different categorical variables). The third, fourth and sixth columns present mean ratios of the estimator's and oracle's ISEs. For the case of $f(y|x, v)$, where no dimension reduction is required/possible, the estimator mimics the oracle very nicely. This outcome is similar to results known for estimation of univariate normal densities; see Efromovich (1999). The performance is worse for experiments with two continuous covariates, but it is still very respectful. Remember that, according to Marron and Wand (1992), one can say that an estimator matches an oracle if its ISE is within 75% of the oracle's ISE. We may conclude that the estimator does mimic the oracle under this criteria as long as we avoid the smallest samples.

The fact that the estimator mimics the oracle relatively well in the case of two continuous covariates implies that it recognizes that $f(y|x, z, v)$ does not depend on z . Frequencies of such correct decisions are shown in columns 5 and 7, and they are respectful keeping in mind the small sample sizes and four-variate dimension of the problem.

6. ACTUARIAL EXAMPLE

Traditionally automobile insurance rates are a function of age, gender, marital status, use of the car, type of car, territory, etc., which are all insurance-type rating variables. Lately some insurers have been using credit score - a non insurance variable - as another rating variable. The insurance industry generally thinks credit score produces a more accurate rating result, and adds something over and above the existing rating variables. Many consumer and regulatory critics think that credit score is inherently discriminatory, and so should not be used. The insurance industry, unlike most other industries, is permitted by law to discriminate - it may not discriminate unfairly. The main study to verify the practice was conducted by the Texas Department of Insurance

(<http://www.tdi.state.tx.us/reports/credit3.html>). Using a multivariate regression analysis, the department concluded that credit score enables insurers to more accurately predict losses. This conclusion is still challenged by many critics.

It would be of interest to repeat that study using a conditional density approach instead of the regression one. Unfortunately, the data is currently not available to the public due to litigation regarding release of the data.

Instead, we shall analyze data volunteered by UT Dallas' students (the author has the university's permission to publish some of the obtained results). Credit score and several standard rating variables were requested. Students were instructed to use the same agency to obtain a free credit score. In their reports many students ignored the request to provide their credit score as well as some other rating variables, but a majority provided information about accident history, grade point average (GPA) and age.

Our aim is to estimate/visualize the conditional density of credit score given accident history and two standard rating variables - GPA and age. This will give us an opportunity to understand how the distribution of credit score, given several standard rating variables, depends on the accident history.

Figure 5 exhibits two (rescaled) scattergrams of credit score versus grade point average (GPA) for two categories of students with none ("No Accident") or at least one auto-accident ("Accident") during last 5 years. Scattergrams are overlaid by regression lines which tell us that credit score improves as GPA increases. It can be also observed that the improvement of credit score occurs faster for the category of students with accident history.

Corresponding estimated conditional densities, presented in Figure 6, give us more information to think about. First of all, conditional densities support "the higher grade, the better credit score" conclusion of the regression analysis. Further, we can conclude that for "No Accident" category of students a regression model with additive heteroscedastic errors, whose volatility decreasing as GPA increases, may be a good first approximation. Further, a normal distribution for regression errors is a good first guess. For "Accident" category of students, regression function per se is not a comprehensive description because the shape of credit score distribution changes dramatically as GPA increases. For larger GPA the distribution resembles the one that we have seen for the "No

Accident” category, only here the volatility of credit score is larger. For smaller GPA the density profile becomes less pronounced and volatility of credit scores dramatically increases. As a result, for the “Accident” category the conditional mean salary (regression line) is no longer the dominant characteristic describing the relationship between credit score and GPA.

Figure 7 shows us the conditional Likelihood Ratio Test Statistic (LRTS) for null hypothesis “No Accident” versus alternative hypothesis “Accident”. Remember that the null hypothesis is rejected when LRTS is small. The interesting part of the shown perspective plot is empty (white) spaces where both conditional densities are zeros. These are the pairs of credit score and GPA that according to the estimates do not occur under either of the hypotheses. What we see is the another appearance of the curse of multidimensionality when almost five hundred files do not provide us with enough information about, say, a student with good grades and poor credit score. We simply do not have such students in our survey, and the conditional density approach highlights this fact. Another remark is that we should be cautious with making any prediction/discrimination for cases near those “white” areas because all estimators perform worse near boundaries.

Overall, our conclusion is that, given GPA, credit score provides important information about accident history.

Now let us add another standard insurance rating variable - age. In the study students were divided into two groups: Mature (more than 22 years old) and Young (at most 22 years old). This, together with the accident history, gives us a new categorical random variable taking 4 values. Scattergrams corresponding to each category are shown in Figure 8. Regression lines indicate that there is a pronounced difference between the category “Accident and Mature” and the three others.

Interestingly, estimated conditional densities for each category, exhibited in Figure 9, do not support that conclusion of the regression analysis. According to the conditional density estimates, an outlier here is the category “Accident and Young” where the conditional density of credit score given GPA changes from almost uniform distribution for smaller GPA to a normal-like distribution for larger GPA. Further, for larger GPA all four conditional densities look almost alike with a normal-like distribution of credit score.

Let us now look more closely on the shape of estimated conditional density for the category “Accident-Mature” highlighted by the regression analysis. For smaller GPA the density does not

correspond to the data shown in the second diagram of Figure 7 where four students with the lowest credit scores have also the smallest GPAs, thus causing the large positive slope, while the estimated density hints that students with smallest GPAs may have excellent credit scores. It is reasonable to conjecture that the conditional density estimate correctly describes the relationship between credit score and GPA for this category of students. Please note that a student with the 4th highest credit score has the 5th smallest GPA, that is, it is possible that an “Accident and Mature” student has a low GPA and a high credit score. Further, if we glance at the data for the category “Accident and Young”, then it is clear that a poor school performance and an accident in the past do not preclude a student from having a perfect credit score. These facts together with extremely small sample size of just 36 students in the category “Accident and Mature” make the shape of conditional density estimate reasonable.

Now let us look at corresponding conditional LRTSs in Figure 10 with the null hypothesis being “No Accident”. The case of mature students is rather straightforward with the only complication that we do not have information about several specific groups of students, including ones with high credit score. The case of young students is more involved because here a rejection region has a more complicated geometry as a function in GPA, the interesting feature of the LRTS is that a high credit score together with a poor school performance raises a red flag about accident history. Please note that this conclusion is supported by visualization of the data shown in Figure 8 which indicates that two highest credit scores among all students were reported by young students with poor school performance and accident history.

We can conclude that the collected data together with the conditional density approach show that even if two classical rating variables age and GPA are given, credit score still gives an Actuary an extra edge in prediction of possible losses.

CONCLUSION

For a regression setting the conditional density of response given explanatory variables (covariates) is the ultimate characteristic describing the relationship between response and explanatory variables. It is a fundamental statistical problem to explore the possibility of a nonparametric estimation of the conditional density for the case of a vector-covariate with some covariates being

categorical. In general, the main obstacle to find a feasible solution is the curse of multidimensionality. There is no way to overcome this curse other than to explore a possibility that the response is conditionally independent of some covariates. This article suggests a two-stage solution of the problem. First, an oracle is suggested that always utilizes the opportunity of dimension reduction and optimal estimation via knowledge of estimated conditional density. Second, an estimator is suggested that mimicks the oracle. It is shown theoretically that the estimator's MISE, up to a constant, matches the oracle's MISE for any sample size. A numerical study shows that the approach is feasible even for sample sizes traditionally challenging for univariate nonparametric estimation. Then a real actuarial example is analyzed using the conditional density approach. It indicates that credit score is a valuable rating variable for a student population even if standard rating variables GPA and age are given.

APPENDIX: PROOFS

In what follows we may skip indexes and sets of summation or integration whenever no confusion occurs. C 's denote finite positive constants that may depend on constants used in assumptions.

Proof of Lemma 1. Denote $X := (U_1, U_2, V)$, $Z := (Y, X)$, $g_{jrst}(Z) := \varphi_j(Y)\psi_r(U_1)\psi_s(U_2)\phi_t(V)$, $\nu := (j, r, s, t)$, and

$$h(Z_l, Z_q) := (m' + 1)^{-2} \sum_{\nu \in B} g_\nu(Z_l)g_\nu(Z_q)p^{-1}(X_l)p^{-1}(X_q),$$

Using this notation we can write

$$\begin{aligned} \hat{\Theta}_{ik} &= L_{ik}^{-1} \frac{2}{n(n-1)} \sum_{1 \leq l < q \leq n} (m' + 1)^{-2} \sum_{\nu \in B_{ik}} g_\nu(Z_l)g_\nu(Z_q)p^{-1}(X_l)p^{-1}(X_q) \\ &= L_{ik}^{-1} \frac{2}{n(n-1)} \sum_{1 \leq l < q \leq n} h(Z_l, Z_q). \end{aligned}$$

The obtained expression for $\hat{\Theta}_{ik}$ hints that it may be beneficial to employ a decoupling technique and then use inequalities of GLZ (Giné, Latole and Zinn 2000). To follow this path, let us introduce an i.i.d. sample Z'_1, \dots, Z'_n which is independent from Z_1, \dots, Z_n and identically distributed, and introduce a decoupled version of $\hat{\Theta}_{ik}$:

$$\hat{\Theta}'_{ik} := L_{ik}^{-1} [n(n-1)]^{-1} \left[\sum_{l,q=1}^n h(Z_l, Z'_q) - \sum_{l=1}^n h(Z_l, Z'_l) \right].$$

Then, according to de la Peña and Montgomery–Smith (1995), to verify Lemma 1 it suffices to establish its validity for the decoupled $\hat{\Theta}'_{ik}$.

Thus, from now on we are exploring $\hat{\Theta}'$ in place of $\hat{\Theta}$ (please note that we are skipping indexes).

To follow GLZ we write

$$\begin{aligned} n(n-1)L\tilde{\Theta}' &= \sum_{l,q=1}^n H(Z_l, Z'_q) + \left[\sum_{l,q=1}^n (E\{h(Z_l, Z'_q)|Z_l\} + E\{h(Z_l, Z'_q)|Z'_q\}) - 2n^2L\Theta \right] \\ &\quad + 2n^2L\Theta - \sum_{l,q=1}^n E\{h(Z_l, Z'_q)\} - \sum_{l=1}^n h(Z_l, Z'_l). \end{aligned}$$

Here

$$H(z, z') := h(z, z') - (E\{h(Z_l, Z'_q)|Z_l = z\} + E\{h(Z_l, Z'_q)|Z'_q = z'\}) + E\{h(Z_l, Z'_q)\}. \quad (\text{A.1})$$

The fact that $E\{h(Z_l, Z'_q)\} = L\Theta$ implies that $H(Z, Z')$ is the completely degenerated and symmetric kernel studied in GLZ. Further, a straightforward algebra yields

$$\begin{aligned} n(n-1)L(\hat{\Theta}' - \Theta) &= \sum_{l,q=1}^n H(Z_l, Z'_l) \\ &\quad + \sum_{l,q=1}^n [(E\{h(Z_l, Z'_q)|Z_l\} - L\Theta) + (E\{h(Z_l, Z'_q)|Z'_q\} - L\Theta)] - \sum_{l=1}^n (h(Z_l, Z'_l) - L\Theta). \end{aligned} \quad (\text{A.2})$$

Using inequalities of GLZ requires knowing upper bounds for several norms of the kernel H .

These bounds are presented below.

Field of science is a lurking variable that , and even the two datasets shown in Figure 1 do not help us to

Lemma A.1. *Suppose that the assumption of Lemma 1 holds. Then*

$$E\{H^2(Z, Z')\} \leq CL, \quad (\text{A.3})$$

$$\|H\|_* := \sup_{\eta_1, \eta_2} \{E\{H(Z, Z')\eta_1(Z)\eta_2(Z')\} : E\{\eta_j^2(Z)\} \leq 1, j = 1, 2\} \leq C, \quad (\text{A.4})$$

$$\sup_z E\{H^2(Z, Z')|Z = z\} \leq CL^{3/2}, \quad (\text{A.5})$$

$$\sup_{z, z'} H(z, z') \leq CL, \quad (\text{A.6})$$

where C 's are generic constants depending only on the constants C_1 and C_2 appearing in the assumption of Lemma 1.

Proof of Lemma A.1. We are verifying the inequalities of Lemma A.1 in turn. Using definition (A.1) and Cauchy inequality yields

$$E\{H^2(Z, Z')\} \leq 2E\{h^2(Z, Z')\} + 8E\{E^2\{h(Z, Z')|Z\}\} + 4(L\Theta)^2. \quad (\text{A.7})$$

For the first term in (A.7) we can write (recall that $Z := (Y, X)$)

$$\begin{aligned} E\{h^2(Z, Z')\} &= E\left\{\left[(m' + 1)^{-2} \sum_{\nu \in B} g_\nu(Z)g_\nu(Z')p^{-1}(X)p^{-1}(X')\right]^2\right\} \\ &= \sum_{\nu, \zeta \in B} E^2\{(m' + 1)^{-2}g_\nu(Z)g_\zeta(Z)p^{-2}(X)\}. \end{aligned} \quad (\text{A.8})$$

Now note that for the considered tensor-product basis the product $g_\nu(z)g_\zeta(z)$ can be written as a finite (no more than $8m'$ terms) linear combination of $g_{\nu'}(z)$ where the first three elements of the vector-index $\nu' := (j', s', r', t')$ are linear combinations of the corresponding elements of ν and ζ . To realize this, remember that for the cosine basis we have $\psi_r(u)\psi_s(u) = 2^{-1/2}[\psi_{r+s}(u) + \psi_{r-s}(u)]$, $r, s \geq 1$, and similar elementary relations hold for the trigonometric basis $\{\varphi_j(y)\}$. Further, $\phi_t(v)\phi_{t'}(v)$, as a function in the categorical variable, can be written using the basis $\{\phi_v, v = 1, \dots, 8\}$. This yields the above-mentioned expansion of $g_\nu(z)g_\zeta(z)$ in no more than $8m'$ additive terms. Denote $E\{(m' + 1)^{-1}g_{\nu'}(Z)p^{-2}(X)\} =: d_{\nu'}$ and then, using the Parseval identity, get $(m' + 1)^{-1} \sum_{t=0}^{m'} \sum_{j,r,s=0}^{\infty} d_{jrst}^2 = \sum_{v=0}^{m'} \int_{[0,1]^3} f^2(y|u_1, u_2, v)p^{-2}(u_1, u_2, v)dydu_1du_2 \leq C < \infty$. This yields

$$E\{h^2(Z, Z')\} \leq CL. \quad (\text{A.9})$$

Further, using Cauchy–Schwarz inequality

$$\begin{aligned} |E\{h(Z, Z')|Z\}| &= |(m' + 1)^{-1} \sum_{\nu \in B} \theta_\nu g_\nu(Z)p^{-1}(X)| \leq (m' + 1)^{-1} \left[\sum_{\nu \in B} \theta_\nu^2 \right]^{1/2} \left[\sum_{\nu \in B} g_\nu^2(Z)p^{-2}(X) \right]^{1/2} \\ &\leq CL^{1/2} \min(1, (L\Theta)^{1/2}). \end{aligned} \quad (\text{A.10})$$

Also, by the Bessel inequality $L\Theta \leq C$, and then combining the obtained results in (A.7) yields (A.3). Now let us verify (A.4). Using (A.1) we can write

$$\|H\|_* \leq \sup_{\eta_1, \eta_2} E\{h(Z, Z')\eta_1(Z)\eta_2(Z')\} + 2 \sup_{\eta_1, \eta_2} E\{E\{h(Z, Z')|Z\}\eta_1(Z)\eta_2(Z')\} + L\Theta. \quad (\text{A.11})$$

To evaluate the first term we begin with a straightforward calculation and then use the Bessel inequality,

$$\begin{aligned}
& \sup_{\eta_1, \eta_2} E\{h(Z, Z')\eta_1(Z)\eta_2(Z')\} = \sup_{\eta_1} \sum_{\nu \in B} E^2\{(m' + 1)^{-1}g_\nu(Z)p^{-1}(X)\eta_1(Z)\} \\
& = \sup_{\eta_1} \sum_{\nu \in B} [(m' + 1)^{-1} \sum_{v=0}^{m'} \int_{[0,1]^3} f(y|u_1, u_2, v)\eta_1(y, u_1, u_2, v)g_\nu(y, u_1, u_2, v)dydu_1du_2]^2 \\
& \leq \sup_{\eta_1} (m' + 1)^{-1} \sum_{v=0}^{m'} \int_{[0,1]^3} f^2(y|u_1, u_2, v)\eta_1^2(y, u_1, u_2, v)dydu_1du_2 \\
& \leq \sup_{y,x} \{f(y|x)p^{-1}(x)\} \sup_{\eta_1} E\{\eta_1^2(Z)\} \leq C. \tag{A.12}
\end{aligned}$$

To consider the second expectation in (A.11) we use $|E\{\eta_2(Z')\}| \leq 1$ and Cauchy-Schwarz inequality. Write,

$$\begin{aligned}
& |E\{E\{h(Z, Z')|Z\}\eta_1(Z)\eta_2(Z')\}| \leq |E\{h(Z, Z')\eta_1(Z)|Z\}| \\
& = |\sum_{\nu \in B} \theta_\nu E\{(m' + 1)^{-1}g_\nu(Z)p^{-1}(X)\eta_1(Z)\}| \leq [\sum_{\nu \in B} \theta_\nu^2]^{1/2} [\sum_{\nu \in B} E^2\{(m' + 1)^{-1}g_\nu(Z)p^{-1}(X)\eta_1(Z)\}]^{1/2} \\
& \leq L^{1/2}\Theta^{1/2}[\sum_{\nu \in B} \kappa_\nu^2]^{1/2} \leq CL^{1/2}\Theta^{1/2} < C. \tag{A.13}
\end{aligned}$$

In the last line we used notation $\kappa_\nu := E\{(m' + 1)^{-1}\eta_1(Z)g_\nu(Z)p^{-1}(X)\}$ and then employed the Bessel inequality to get $\sum_{\nu \in B} \kappa_\nu^2 \leq \sum_{v=0}^{m'} (m' + 1)^{-1} \int_{[0,1]^3} f^2(y|u_1, u_2, v)\eta_1^2(y, u_1, u_2, v)dydu_1du_2 \leq \sup_{y,x} \{f(y|x)p^{-1}(x)\}E\{\eta_1^2\} \leq C$. Combining (A.11)–(A.13) yields (A.4).

To check (A.5) we write, using (A.1), Cauchy inequality and a straightforward algebra,

$$\begin{aligned}
& E\{H^2(Z, Z')|Z = z\} = E\{H^2(z, Z')\} \\
& \leq 2E\{h^2(z, Z')\} + 4E^2\{h(z, Z')\} + 8E\{E^2\{h(Z, Z')|Z'\}\} + 8L^2\Theta^2 \\
& \leq 2E\{[\sum_{\nu \in B} (m' + 1)^{-2}g_\nu(z)g_\nu(Z')p^{-1}(x)p^{-1}(X')]\}^2 + 4[\sum_{\nu \in B} (m' + 1)^{-1}g_\nu(z)p^{-1}(x)\theta_\nu]^2 \\
& \quad + 8E\{[\sum_{\nu \in B} (m' + 1)^{-1}\theta_\nu g_\nu(Z')p^{-1}(X')]\}^2 + 8L^2\Theta^2. \tag{A.14}
\end{aligned}$$

Now we are evaluating terms in (A.14) in turn. Recall that all elements of the considered tensor-product basis are uniformly bounded and $p^{-1}(x)$ is also bounded. Using this together with Cauchy-Schwarz inequality allows us to write,

$$E\{[\sum_{\nu \in B} (m' + 1)^{-2}g_\nu(z)p^{-1}(x)g_\nu(Z')p^{-1}(X')]\}^2$$

$$\begin{aligned}
&= \sum_{\nu, \zeta \in B} g_\nu(z) g_\zeta(z) p^{-2}(x) E\{(m' + 1)^{-4} g_\nu(Z') g_\zeta(Z') p^{-2}(X')\} \\
&\leq C \sum_{\nu, \zeta \in B} |E\{(m' + 1)^{-4} g_\nu(Z') g_\zeta(Z') p^{-2}(X')\}| \\
&\leq CL \left[\sum_{\nu, \zeta \in B} E^2\{(m' + 1)^{-4} g_\nu(Z') g_\zeta(Z') p^{-2}(X')\} \right]^{1/2} \leq CL^{3/2}.
\end{aligned}$$

In the last inequality we used (A.8)–(A.9). To evaluate the second term in (A.14) we use Cauchy-Schwarz inequality and get

$$\sup_{z, x} \left[\sum_{\nu \in B} g_\nu(z) p^{-1}(x) \theta_\nu \right]^2 \leq \sum_{\nu \in B} \theta_\nu^2 \sup_{z, x} \sum_{\nu \in B} g_\nu^2(z) p^{-2}(x) \leq CL.$$

The last inequality also yields $E\{[\sum_{\nu \in B} (m' + 1)^{-1} \theta_\nu g_\nu(Z') p^{-1}(X')]^2\} \leq CL$. Combining the obtained results in (A.14) verifies (A.5).

Finally, inequality (A.6) is based on the boundness of $g_\nu(z) p^{-1}(x)$. Lemma A.1 is proved.

Let us continue our proof of (23). To evaluate the expectation of the squared first term in (A.2) we use Corollary 3.4 of GLZ which, together with Lemma A.1, yields

$$\begin{aligned}
E \left| \sum_{l, q=1}^n H(Z_l Z'_q) \right|^2 &\leq C [n^2 E\{H^2(Z, Z')\} + n^2 \|H\|_*^2 + n \sup_z E\{H^2(z, Z')\} + (\sup_{z, z'} H(z, z'))^2] \\
&\leq C [n^2 L + nL^{3/2} + L^2] \leq CLn^2.
\end{aligned} \tag{A.15}$$

In the last inequality we used our assumption $L < C^*n$.

Further, to evaluate the expectation of the squared second term in (A.2), we begin with the following relations,

$$\begin{aligned}
&E \left[\sum_{l, q=1}^n [E\{h(Z_l, Z'_q) | Z_l\} - L\Theta] \right]^2 = E \left[\sum_{l, q=1}^n \left[\sum_{\nu \in B} ((m' + 1)^{-1} g_\nu(Z_l) p^{-1}(X_l) \theta_\nu - \theta_\nu^2) \right] \right]^2 \\
&= n^2 n E \left[\sum_{\nu \in B} \theta_\nu ((m' + 1)^{-1} g_\nu(Z) p^{-1}(X) - \theta_\nu) \right]^2 \leq n^3 \sum_{\nu, \zeta \in B} \theta_\nu \theta_\zeta E\{(m' + 1)^{-2} g_\nu(Z) g_\zeta(Z) p^{-2}(X)\} \\
&\leq n^3 \sum_{\nu \in B} \theta_\nu^2 \sum_{\zeta \in B} |E\{(m' + 1)^{-2} g_\nu(Z) g_\zeta(Z) p^{-2}(X)\}|.
\end{aligned} \tag{A.16}$$

In the last inequality we used Cauchy inequality for $|\theta_\nu \theta_\zeta|$. Then, similarly to the technique used in the paragraph between lines (A.8) and (A.9), we can show that $\max_{\nu} \sum_{\zeta \in B} |E\{(m' + 1)^{-2} g_\nu(Z) g_\zeta(Z) p^{-2}(X)\}| \leq CL^{1/2}$. Using this in (A.16) yields

$$E \left[\sum_{l, q=1}^n [E\{h(Z_l, Z'_q) | Z_l\} - L\Theta] \right]^2 \leq Cn^3 L^{3/2} \Theta. \tag{A.17}$$

To evaluate the expectation of the last squared sum in (A.2) we use (A.9) and get

$$E\left[\sum_{l=1}^n (h(Z_l, Z'_l) - L\Theta)^2\right] \leq nE\{h^2(Z_l, Z'_l)\} \leq CnL. \quad (\text{A.18})$$

Using (A.15), (A.17) and (A.18) together with (A.2) allows us to conclude that

$$E(\tilde{\Theta}' - \Theta)^2 \leq CL^{-2}n^{-4}[n^2L + n^3L^{3/2}\Theta + nL] \leq C_3L^{-1}n^{-1}[L^{1/2}\Theta + n^{-1}],$$

where C_3 depends only on (C_1, C_2, C^*) . This proves the moment inequality (23) of Lemma 1.

Now we are verifying the exponential inequality (24). According to Corollary 3.4 in GLZ and Lemma A.1, for any positive constant a the following inequality holds,

$$\Pr\left\{\left|\sum_{l,q=1}^n H(Z_l, Z'_q)\right| \geq a\right\} \leq C \exp\left\{-C^{-1} \min\left[\frac{a^2}{n^2L}, \frac{a}{n}, \frac{a^{2/3}}{[nL^{3/2}]^{1/3}}, \frac{a^{1/2}}{L^{1/2}}\right]\right\}. \quad (\text{A.19})$$

This result allows us to get an exponential inequality for the first term in (A.2). For the next term we can write

$$\sum_{l,q=1}^n [E\{h(Z_l, Z'_q)|Z_l\} - L\Theta] = n \sum_{l=1}^n \sum_{\nu \in B} \theta_\nu ((m'+1)^{-1}g_\nu(Z_l)p^{-1}(X_l) - \theta_\nu). \quad (\text{A.20})$$

Remember that according to the Bernstein inequality, if W_1, W_2, \dots, W_n are i.i.d., $|W_l| \leq M < \infty$ a.e., $E\{W_l\} = 0$ and $\text{Var}(W_l) = \sigma^2 < \infty$ then for any positive a

$$\max(\Pr\{\sum_{l=1}^n W_l < -a\}, \Pr\{\sum_{l=1}^n W_l > a\}) \leq \exp\left\{-\frac{a^2}{2n\sigma^2 + (2/3)Ma}\right\}. \quad (\text{A.21})$$

To use the Bernstein inequality in (A.20), we use notation $W_l := \sum_{\nu \in B} \theta_\nu ((m'+1)^{-1}g_\nu(Z_l)p^{-1}(X_l) - \theta_\nu)$, and then note that $E\{W_l\} = 0$, $|W_l| \leq [\sum_{\nu \in B} \theta_\nu^2 \sum_{\nu \in B} ((m'+1)^{-1}g_\nu(Z_l)p^{-1}(X_l) - \theta_\nu)^2]^{1/2} \leq CL\Theta^{1/2}$. This, together with (A.16) and (A.17), yields

$$\begin{aligned} \text{Var}(W_l) &= E\{W_l^2\} \\ &= E\left\{\sum_{\nu, \zeta \in B} \theta_\nu \theta_\zeta ((m'+1)^{-1}g_\nu(Z_l)p^{-1}(X_l) - \theta_\nu) ((m'+1)^{-1}g_\zeta(Z_l)p^{-1}(X_l) - \theta_\zeta)\right\} \leq CL^{3/2}\Theta. \end{aligned}$$

Then we are using the Bernstein inequality and get:

$$\begin{aligned} \Pr\left(\left|\sum_{l,q=1}^n [E\{h(Z_l, Z'_q)|Z_l\} - L\Theta]\right| > a\right) &= \Pr\left(n \left|\sum_{l=1}^n W_l\right| > a\right) \\ &\leq 2 \exp\left\{-\frac{a^2 n^{-2}}{C(nL^{3/2}\Theta + L\Theta^{1/2}an^{-1})}\right\}. \end{aligned} \quad (\text{A.22})$$

Finally, let us use the Bernstein inequality for the analysis of the last term in (A.2). Define $W'_l := h(Z_l, Z'_l) - L\Theta$. Immediately we get $E\{W'_l\} = 0$, $|W'_l| \leq CL$, and also $\text{Var}(W'_l) \leq CL$ according to (A.9). This yields

$$\Pr\left(\left|\sum_{l=1}^n (h(Z'_l) - L\Theta)\right| > a\right) \leq 2 \exp\left\{-\frac{a^2}{CL(n+a)}\right\}. \quad (\text{A.23})$$

In particular, for $a = \kappa tnL$ the obtained results, together with (A.2), (A.19), (A.22) and (A.23), imply that

$$\begin{aligned} & \Pr\{n(n-1)L|\tilde{\Theta}' - \Theta| > \kappa tnL\} \\ & \leq C \left[\exp\left\{-C^{-1} \min\left[\frac{\kappa^2 t^2 n^2 L^2}{n^2 L}, \frac{\kappa tnL}{n}, \frac{\kappa^{2/3} t^{2/3} n^{2/3} L^{2/3}}{n^{1/3} L^{1/2}}, \frac{\kappa^{1/2} t^{1/2} n^{1/2} L^{1/2}}{L^{1/2}}\right]\right\} \right. \\ & \quad \left. + \exp\left\{-C^{-1} \frac{\kappa^2 t^2 n^2 L^2 n^{-2}}{nL^{3/2}\Theta + L\Theta^{1/2}\kappa tnLn^{-1}}\right\} + \exp\left\{-C^{-1} \frac{\kappa^2 t^2 n^2 L^2}{Ln(1 + \kappa tL)}\right\} \right] \\ & \leq C \left[\exp\left\{-C^{-1} \min[\kappa^2 t^2 L, \kappa tL, \kappa^{2/3} t^{2/3} n^{1/3} L^{1/6}, (\kappa tn)^{1/2}]\right\} \right. \\ & \quad \left. + \exp\left\{-C^{-1} \frac{\kappa^2 t^2 L^{1/2}}{n\Theta + \kappa t(L\Theta)^{1/2}}\right\} + \exp\left\{-C^{-1} \frac{\kappa^2 t^2 Ln}{1 + \kappa tL}\right\} \right]. \quad (\text{A.24}) \end{aligned}$$

Lemma 1 is established.

Proof of Theorems 1 and 2. The Parseval identity allows us to evaluate the estimator's risk in the left side of (17) via considering a particular block B_{ik} . Recall our notation $\nu := (j, r, s, t)$. The Cauchy–Schwarz inequality yields for a block B_{ik} ,

$$\begin{aligned} & E\left\{\sum_{\nu \in B_{ik}} (\tilde{\mu}_{ik}\tilde{\theta}_\nu - \theta_\nu)^2\right\} = E\left\{\sum_{\nu \in B_{ik}} [(\mu_{ik}\tilde{\theta}_\nu - \theta_\nu) + (\tilde{\mu}_{ik} - \mu_{ik})\tilde{\theta}_\nu]^2\right\} \\ & \leq (1+c)E\left\{\sum_{\nu \in B_{ik}} (\mu_{ik}\tilde{\theta}_\nu - \theta_\nu)^2\right\} + (1+c^{-1})E\left\{(\tilde{\mu}_{ik} - \mu_{ik})^2 \sum_{\nu \in B_{ik}} \tilde{\theta}_\nu^2\right\}. \quad (\text{A.25}) \end{aligned}$$

Here $c \in (0, 1)$. Now we need to establish two basic statistical properties of the estimate $\tilde{\theta}_\nu$.

Lemma A.2. *Suppose that Assumption 1 holds. Then the estimate $\tilde{\theta}_\nu$, defined in (6), is an unbiased estimate of θ_ν meaning that*

$$E\{\tilde{\theta}_\nu\} = \theta_\nu. \quad (\text{A.26})$$

If additionally (15) holds then variance of the estimate satisfies the following relation

$$\sum_{j=2q-1}^{2q} \text{Var}(\tilde{\theta}_{jrst}) = n^{-1}[2d_{rst} - (\theta_{(2q-1)rst}^2 + \theta_{(2q)rst}^2)], \quad q = 1, 2, \dots \quad (\text{A.27})$$

Here

$$d_{rst} := d_{rst}(p) := E\{(m' + 1)^{-2} \psi_r^2(U_1) \psi_s^2(U_2) \phi_t^2(V) p^{-2}(U_1, U_2, V)\} \quad (\text{A.28})$$

have the following properties:

$$d_{rst} \rightarrow d_t := E\{(m' + 1)^{-2} \phi_t^2(V) p^{-2}(U_1, U_2, V)\} \text{ as } \min(r, s) \rightarrow \infty, \quad (\text{A.29})$$

and for the case of the uniform design density $p(u_1, u_2, v) \equiv (m' + 1)^{-1}$

$$d_{rst} \equiv 1. \quad (\text{A.30})$$

Further, it follows from (A.27) that

$$D_{ik} - \Theta_{ik} > 0 \quad (\text{A.31})$$

where

$$D_{ik} := L_{ik}^{-1} \sum_{(j,r,s,t) \in B_{ik}} d_{rst}. \quad (\text{A.32})$$

Further, the oracle's shrinkage coefficient (7) can be written as

$$\mu_{ik} = \frac{\Theta_{ik}}{\Theta_{ik} + n^{-1}(D_{ik} - \Theta_{ik})}. \quad (\text{A.33})$$

In particular, for the case of the uniform design

$$\mu_{ik} = \frac{\Theta_{ik}}{\Theta_{ik} + n^{-1}(1 - \Theta_{ik})}. \quad (\text{A.34})$$

Proof of Lemma A.2. Remember (1), (6) and write

$$\begin{aligned} E\{\tilde{\theta}_{jrst}\} &= \sum_{v=0}^{m'} (m' + 1)^{-1} \int_{[0,1]^3} f(y|u_1, u_2, v) p(u_1, u_2, v) \\ &\quad \times \varphi_j(y) \psi_r(u_1) \psi_s(u_2) \phi_t(v) p^{-1}(u_1, u_2, v) dy du_1 du_2 = \theta_{jrst}. \end{aligned}$$

This verifies (A.26). Further, for any positive integer q

$$\begin{aligned} \sum_{j=2q-1}^{2q} \text{Var}(\tilde{\theta}_{jrst}) &= n^{-1} \sum_{j=2q-1}^{2q} E\{[(m' + 1)^{-1} \varphi_j(Y) \psi_r(U_1) \psi_s(U_2) \phi_t(V) p^{-1}(U_1, U_2, V) - \theta_{jrst}]^2\} \\ &= n^{-1} [E\{(\varphi_{2q-1}^2(y) + \varphi_{2q}^2(y)) (m' + 1)^{-2} \psi_r^2(U_1) \psi_s^2(U_2) \phi_t^2(V) p^{-2}(U_1, U_2, V)\} - (\theta_{(2q-1)rst}^2 + \theta_{(2q)rst}^2)]. \end{aligned} \quad (\text{A.35})$$

For the considered bases: $\varphi_{2q-1}^2(y) + \varphi_{2q}^2(y) = 2$, $y \in [0, 1]$; $\psi_r^2(u) = 1 + 2^{-1/2}\psi_{2r}(u)$, $r > 0$ and $\psi_0^2(u) = 1$; $(m' + 1)^{-1} \sum_{v=0}^{m'} \phi_t^2(v) = 1$, $t \in \{0, 1, \dots, m'\}$. Recall that $\int_0^1 f(y|u_1, u_2, v)dy = 1$, and then a simple algebra yields (A.27). Relation (A.29) immediately follows from (A.27), the above-mentioned properties of the bases, and the Bessel inequality. Relations (A.33)-(A.34) are based on (A.27), (A.32) and an elementary algebra. Lemma A.2 is proved.

Returning to (A.25), let us note that its first expectation is a part of the oracle's risk corresponding to the block B_{ik} ; let us calculate it because this will verify Theorem 2. Recall (7), Lemma A.2 and our convention that indexes may be skipped whenever no confusion occurs. Write,

$$\begin{aligned} \sum_{\nu \in B} E(\mu\tilde{\theta}_\nu - \theta_\nu)^2 &= \sum_{\nu \in B} E[\mu(\tilde{\theta}_\nu - \theta_\nu) - (1 - \mu)\theta_\nu]^2 \\ &= L\mu^2 n^{-1}(D - \Theta) + (1 - \mu)^2 L\Theta \\ &= L \frac{\Theta^2 n^{-1}(D - \Theta)}{[\Theta + n^{-1}(D - \Theta)]^2} + L \frac{n^{-2}(D - \Theta)^2 \Theta}{[\Theta + n^{-1}(D - \Theta)]^2} = n^{-1} L\mu(D - \Theta). \end{aligned} \quad (\text{A.36})$$

The last equality, the Parseval identity and Lemma A.2 verify Theorem 2.

Now we are considering the second expectation in (A.25). Write,

$$\begin{aligned} E\{(\tilde{\mu} - \mu)^2 \sum_{\nu \in B} \tilde{\theta}_\nu^2\} &= E\{(\tilde{\mu} - \mu)^2 \sum_{\nu \in B} \tilde{\theta}_\nu^2 I(\hat{\Theta} > tn^{-1})\} + E\{(\tilde{\mu} - \mu)^2 \sum_{\nu \in B} \tilde{\theta}_\nu^2 I(\hat{\Theta} \leq tn^{-1})\} \\ &=: E\{A_1\} + E\{A_2\}. \end{aligned} \quad (\text{A.37})$$

To make our next step, let us remember our notation: $\nu := (j, r, s, t)$, $z := (y, u_1, u_2, v)$, $x := (u_1, u_2, v)$, $g_\nu(z) := \varphi_j(y)\psi_r(u_1)\psi_s(u_2)\phi_t(v)$, $\tilde{\theta}_\nu := n^{-1}(m' + 1)^{-1} \sum_{l=1}^n g_\nu(Z_l)p^{-1}(X_l)$, $h_{ik}(z_l, z_q) = \sum_{\nu \in B_{ik}} (m' + 1)^{-2} g_\nu(z_l)g_\nu(z_q)p^{-1}(z_l)p^{-1}(z_q)$. Also remember that

$$\tilde{\Theta} := L^{-1} \sum_{\nu \in B} \tilde{\theta}_\nu^2$$

is a biased estimate of Θ and

$$\hat{\Theta} := L^{-1}[n(n-1)]^{-1} \sum_{l \neq q=1}^n h(Z_l, Z_q)$$

is an unbiased estimate of Θ . These two estimates are used in $\tilde{\mu}$ (recall (10)), and our first step is to understand how they are related. Write

$$\tilde{\Theta} = L^{-1} \sum_{\nu \in B} \tilde{\theta}_\nu^2 = L^{-1} n^{-2} \sum_{\nu \in B} \sum_{l, q=1}^n (m' + 1)^{-2} g_\nu(Z_l)g_\nu(Z_q)p^{-1}(X_l)p^{-1}(X_q)$$

$$\begin{aligned}
&= \hat{\Theta} + (n^{-2} - [n(n-1)]^{-1})n(n-1)\hat{\Theta} + n^{-2}L^{-1} \sum_{\nu \in B} \sum_{l=1}^n (m' + 1)^{-2} g_\nu^2(Z_l) p^{-2}(X_l) \\
&= \hat{\Theta}(1 - n^{-1}) + n^{-2}L^{-1} \sum_{l=1}^n h(Z_l, Z_l). \tag{A.38}
\end{aligned}$$

Using (15), (A.28) and (A.32) yields

$$\begin{aligned}
n^{-1}L^{-1} \sum_{l=1}^n E\{h(Z_l, Z_l)\} &= L^{-1} \sum_{\nu \in B} E\{(m' + 1)^{-2} g_\nu^2(Z) p^{-2}(X)\} \\
&= L^{-1} \sum_{\nu \in B} E\{(m' + 1)^{-2} \varphi_j^2(Y) \psi_r^2(U_1) \psi_s^2(U_2) \phi_i^2(V) p^{-2}(X)\} = D.
\end{aligned}$$

Then, if we denote

$$\hat{D} := n^{-1}L^{-1} \sum_{l=1}^n h(Z_l, Z_l), \tag{A.39}$$

we get $E\{\hat{D}\} = D$, and this allows us to rewrite (A.38) as

$$\tilde{\Theta} = \hat{\Theta}(1 - n^{-1}) + n^{-1}\hat{D} = \hat{\Theta} + n^{-1}(\hat{D} - \hat{\Theta}). \tag{A.40}$$

Further, recall (10), (A.33) and note that for $\hat{\Theta} > tn^{-1}$

$$\begin{aligned}
\tilde{\mu} - \mu &= \frac{\hat{\Theta}}{\hat{\Theta} + n^{-1}(\hat{D} - \hat{\Theta})} - \frac{\Theta}{\Theta + n^{-1}(D - \Theta)} \\
&= \frac{n^{-1}[\hat{\Theta}(D - \Theta) - \Theta(\hat{D} - \hat{\Theta})]}{(\hat{\Theta} + n^{-1}(\hat{D} - \hat{\Theta}))(\Theta + n^{-1}(D - \Theta))} = \frac{n^{-1}[(\hat{\Theta} - \Theta)D + \Theta(D - \hat{D})]}{(\hat{\Theta} + n^{-1}(\hat{D} - \hat{\Theta}))(\Theta + n^{-1}(D - \Theta))}.
\end{aligned}$$

Using this expression, together with (A.33), (A.40), notation $\sum_{\nu \in B} \tilde{\theta}_\nu^2 = L\tilde{\Theta}$ and our assumption $n > 3$, allows us to evaluate the term A_1 defined in (A.37):

$$\begin{aligned}
A_1 &\leq \frac{2Ln^{-2}(\hat{\Theta} - \Theta)^2 D^2 + 2Ln^{-2}\Theta^2(D - \hat{D})^2}{(\hat{\Theta} + n^{-1}(\hat{D} - \hat{\Theta}))(\Theta + n^{-1}(D - \Theta))^2} I(\hat{\Theta} > tn^{-1}) \\
&\leq 3Lt^{-1}n^{-1}D^2(\hat{\Theta} - \Theta)^2[\Theta + n^{-1}(D - \Theta)]^{-2} I(\Theta > qtn^{-1}) \\
&\quad + 3Ln^{-2}D^2(\hat{\Theta} - \Theta)[\Theta + n^{-1}(D - \Theta)]^{-2} I(\hat{\Theta} - \Theta > (1 - q)tn^{-1}) I(\Theta \leq qtn^{-1}) \\
&\quad + 3L\mu n^{-1}[t^{-1}\mu(D - \hat{D})^2] =: A_{11} + A_{12} + A_{13}. \tag{A.41}
\end{aligned}$$

Here $q \in (0, 1)$. Now we are evaluating the expectation of the last three terms in turn. Using Lemma 1, (19) and a relation

$$(L^{1/2}\Theta + n^{-1})\Theta^{-1}I(\Theta > qtn^{-1}) = (L^{1/2} + n^{-1}\Theta^{-1})I(\Theta > qtn^{-1}) \leq (L^{1/2} + q^{-1}t^{-1})I(\Theta > qtn^{-1})$$

we get,

$$\begin{aligned} E\{A_{11}\} &\leq \frac{3\mu Lt^{-1}n^{-1}D^2C_3L^{-1}n^{-1}(L^{1/2}\Theta + n^{-1})}{[\Theta + n^{-1}(D - \Theta)]\Theta} I(\Theta > qtn^{-1}) \\ &\leq L\mu n^{-1}[3C_3t^{-1}\max(1, (D - \Theta)^{-1})D^2L^{-1}[L^{1/2} + q^{-1}t^{-1}]I(\Theta > qtn^{-1})]. \end{aligned}$$

Similarly

$$\begin{aligned} E\{A_{12}\} &\leq 3Ln^{-2}D^2E^{1/2}\{(\hat{\Theta} - \Theta)^2\}P^{1/2}(\hat{\Theta} - \Theta > (1 - q)tn^{-1})[\Theta + n^{-1}(D - \Theta)]^{-2}I(\Theta \leq qtn^{-1}) \\ &\leq 3\max(1, (D - \Theta)^{-2})D^2Ln^{-2}C_3^{1/2}L^{-1/2}n^{-1/2}(L^{1/2}\Theta + n^{-1})^{1/2}[\Theta + n^{-1}]^{-2} \\ &\quad \times P^{1/2}(\hat{\Theta} - \Theta > (1 - q)tn^{-1})I(\Theta \leq qtn^{-1}) \\ &\leq n^{-1}\max(1, (D - \Theta)^{-2})D^2C_3^{1/2}L^{3/4}P^{1/2}(\hat{\Theta} - \Theta > (1 - q)tn^{-1})I(\Theta \leq qtn^{-1}). \end{aligned}$$

Plainly

$$E\{(\hat{D} - D)^2\} \leq C_4n^{-1}, \quad (\text{A.42})$$

where C_4 depends only on the constant C_2 introduced in Assumption 1. This yields

$$E\{A_{13}\} \leq L\mu n^{-1}[3C_4t^{-1}\mu n^{-1}]I(\Theta > qtn^{-1}) + L\mu n^{-1}[3C_4n^{-1}]I(\Theta \leq qtn^{-1}).$$

(In what follows C_i 's are finite constants depending only on constants in the assumption.) Combining the obtained results, we get

$$\begin{aligned} E\{A_1\} &\leq L\mu n^{-1}\{[C_5t^{-1}L^{-1}(L^{1/2} + q^{-1}t^{-1})]I(\Theta > qtn^{-1}) + 3C_4n^{-1}I(\Theta \leq qtn^{-1})\} \\ &\quad + n^{-1}C_5L^{3/4}P^{1/2}(\hat{\Theta} - \Theta > (1 - q)tn^{-1})I(\Theta \leq qtn^{-1}). \end{aligned} \quad (\text{A.43})$$

Further, let us evaluate the expectation of A_2 defined in (A.37). In A_2 we have $\tilde{\mu} = 0$. Then, using (12), (A.40), boundness of \hat{D} , (A.33) and (23) we get

$$\begin{aligned} E\{A_2\} &= \mu^2LE\{\tilde{\Theta}I(\hat{\Theta} \leq tn^{-1})\} \\ &= \mu^2LE\{[\hat{\Theta}(1 - n^{-1}) + n^{-1}\hat{D}]I(\hat{\Theta} \leq tn^{-1})\} \\ &\leq \mu^2Ln^{-1}t + \mu^2Ln^{-1}E\{\hat{D}I(\hat{\Theta} \leq tn^{-1})[I(\Theta > 2tn^{-1}) + I(\Theta \leq 2tn^{-1})]\} \\ &\leq \mu Ln^{-1}(\mu t) + \mu Ln^{-1}E\{\mu\hat{D}\frac{(\Theta - \hat{\Theta})^2}{(1/4)\Theta^2}I(\Theta > 2tn^{-1})\} + \mu Ln^{-1}\frac{2t}{2t + D - \Theta} \end{aligned}$$

$$\leq \mu L n^{-1} C_6 [L^{-1/2} t^{-1} + t \max(1, (D - \Theta)^{-1})].$$

Returning to (A.37) we conclude that

$$\begin{aligned} E\{(\tilde{\mu} - \mu)^2 \sum_{\nu \in B} \tilde{\theta}_\nu^2\} &\leq L \mu n^{-1} C_7 [L^{-1/2} t^{-1} + \max(1, (D - \Theta)^{-1}) t + t^{-2} L^{-1} q^{-1} + n^{-1}] \\ &\quad + C_5 n^{-1} L^{3/4} P^{1/2} (\hat{\Theta} - \Theta > (1 - q) t n^{-1}) I(\Theta \leq q t n^{-1}). \end{aligned}$$

Using inequality (24) with $\kappa = 1 - q$ and $q = \min(1/2, t^{-1} L^{-1/2})$ we continue our evaluation,

$$\begin{aligned} E\{(\tilde{\mu} - \mu)^2 \sum_{\nu \in B} \tilde{\theta}_\nu^2\} &\leq L \mu n^{-1} C_8 [L^{-1/2} t^{-1} + t \max(1, (D - \Theta)^{-1}) + n^{-1}] \\ &\quad + C_9 n^{-1} L^{3/4} \left[\exp\{-C_9^{-1} \min(t^2 L, tL, t^{2/3} n^{1/3} L^{1/6}, (tn)^{1/2})\} \right. \\ &\quad \left. + \exp\{-C_9^{-1} t^2 L [\min(1, tL^{1/2}) + n^{-1/2} L^{3/4} t \min(1, t^{1/2} L^{1/4})]\}^{-1} \right. \\ &\quad \left. + \exp\{-C_9^{-1} t^2 L n (1 + tL)^{-1}\} \right]. \end{aligned} \quad (\text{A.44})$$

Using (14) and Assumption 2 allows us to simplify (A.44),

$$E\{(\tilde{\mu}_{ik} - \mu_{ik})^2 \sum_{\nu \in B_{ik}} \tilde{\theta}_\nu^2\} \leq n^{-1} L_{ik} \mu_{ik} (\delta'_k + C_8 n^{-1}) + C_{10} n^{-1} L_{ik}^{3/4} \exp\{-C_{10}^{-1} t_{ik}^2 L_{ik}\}, \quad (\text{A.45})$$

where $\delta'_k \rightarrow 0$ as $k \rightarrow \infty$.

Using (A.36) and (A.45) in (A.25) we conclude that for any $c \in (0, 1)$

$$\begin{aligned} E\left\{ \sum_{\nu \in B_{ik}} (\tilde{\mu}_{ik} \tilde{\theta}_\nu - \theta_\nu)^2 \right\} &\leq (1 + c) n^{-1} L_{ik} \mu_{ik} (D_{ik} - \Theta_{ik}) [1 + (\delta'_k + C_8 n^{-1}) (D_{ik} - \Theta_{ik})^{-1}] \\ &\quad + n^{-1} (1 + c^{-1}) C_{10} L_{ik}^{3/4} \exp\{-C_{10}^{-1} t_{ik}^2 L_{ik}\}. \end{aligned} \quad (\text{A.46})$$

Then the Parseval identity, together with (A.46), Assumption 2 and already proved Theorem 2, implies that

$$\begin{aligned} &(m' + 1)^{-1} E\left\{ \sum_{v=0}^{m'} \int_{[0,1]^3} (\tilde{f}(y|u_1, u_2, v) - f(y|u_1, u_2, v))^2 dy du_1 du_2 \right\} \\ &\leq (1 + c)(1 + \delta_n)(m' + 1)^{-1} E\left\{ \sum_{v=0}^{m'} \int_{[0,1]^3} (\tilde{f}^*(y|u_1, u_2, v) - f(y|u_1, u_2, v))^2 dy du_1 du_2 \right\} + (1 + c^{-1}) C_* n^{-1}, \end{aligned}$$

where $C_* < \infty$, $\delta_n \leq C < \infty$, and if the estimated conditional density f is nonparametric then $\delta_n \rightarrow 0$ as $n \rightarrow \infty$. Theorem 1 is proved, and recall that Theorem 2 has been proved earlier.

References

- Brown, L.D., Low, M.G., and Zhao.L.H. (1997), “Superefficiency in Nonparametric Function Estimation,” *The Annals of Statistics*, 25, 2607–2625.
- de la Peña, V. and Montgomery-Smith, S. (1995), “Decoupling inequalities for the tail probabilities of multivariate U-statistics,” *The Annals of Probability*, 23, 806–816.
- Doksum, K., and Samarov, A. (1995), “Nonparametric Estimation of Global Functionals and a Measure of the Explanatory Power of Covariates in Regression,” *The Annals of Statistics*, 23, 1443–1473.
- Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory and Applications*, New York: Springer.
- Efromovich, S. (2000), “On Sharp Adaptive Estimation of Multivariate Curves,” *Mathematical Methods of Statistics*, 9, 117–139.
- Efromovich, S. (2001), “Density Estimation Under Random Censorship and Order Restrictions: From Asymptotic to Small Sample Sizes,” *Journal of the American Statistical Association*, 94, 667–685.
- Efromovich, S. (2004), “Oracle Inequality for Efromovich-Pinsker Blockwise Estimator,” *Methodology and Computing in Applied Probability*, 6, 303–322.
- Efromovich, S. (2007), “Conditional Density Estimation in a Regression Setting,” *Annals of Statistics*, 35, 2504–2535.
- Eubank, R.L. (1999), *Spline Smoothing and Nonparametric Regressions*, 2nd. ed. New York: Marcel and Dekker.
- Fan, J. and Yim, T.H. (2004), “A Cross-Validation method for Estimating conditional Densities,” *Biometrika*, 91, 819–834.
- Giné, E., Latala, R. and Zinn, J. (2000), “Exponential and Moment Inequalities for U-statistics,” *High Dimensional Probability 2, Progress in Probability*, 47, 13–38.
- Hall, P. (1983). “Orthogonal Series Methods for Both Qualitative and Quantitative Data,” *Annals of Statistics*, 11, 1004–1007.
- Hall, P., Racine, J. and Li, Q. (2004). “Cross-Validation and the Estimation of Conditional Probability Densities,” *Journal of the American Statistical Association*, 99, 1015–1026.
- Hall, P. and Yao, Q. (2005). “Approximating Conditional Distribution Functions Using Dimension Reduction,” *Annals of Statistics*, 33, 1404–1422.
- Hart, J.D. (1997), *Nonparametric Smoothing and Lack-Of-Fit Tests*, New York: Springer.
- Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.K. (1996). “Estimating and Visualizing Conditional Densities,” *Journal of Computational and Graphics Statistics*, 5, 315–336.

- Hyndman, R.J. and Yao, Q. (2002). "Nonparametric Estimation and Symmetry Tests for Conditional Density Functions," *Nonparametric Statistics*, 14, 259-278.
- Johnstone, I. (1998). *Function Estimation in Gaussian Noise*, Draft of Monograph, Stanford University, California.
- Marron, J.S. and Wand, M.P.(1992). Exact mean integrated error. *Annals of Statistics*, 20, 712–736.
- Moore, D.S. and McGabe, G.P. (2003). *Introduction to the Practice of Statistics*, W.H.Freeman and Co: New York.
- Samarov, A. (1993), "Exploring Regression Structure Using Nonparametric Functional Estimation," *Journal of the American Statistical Association*, 88, 836-849.
- Samarov, A. and Tsybakov, A. (2007), "Aggregation of density estimators and dimension reduction". In *Advances in Statistical Modeling and Inference*. Essays in Honor of Kjell A. Doksum, V. Nair, ed., 233-251.
- Simonoff, J.S. (1996), *Smoothing Methods in Statistics*, New York: Springer.
- Yang, Y. (2000), "Mixing strategies for density estimation," *Annals of Statistics*, 28, 75-87.
- Wasserman, L. (2005). *All of Nonparametric Statistics*, Springer: New York.

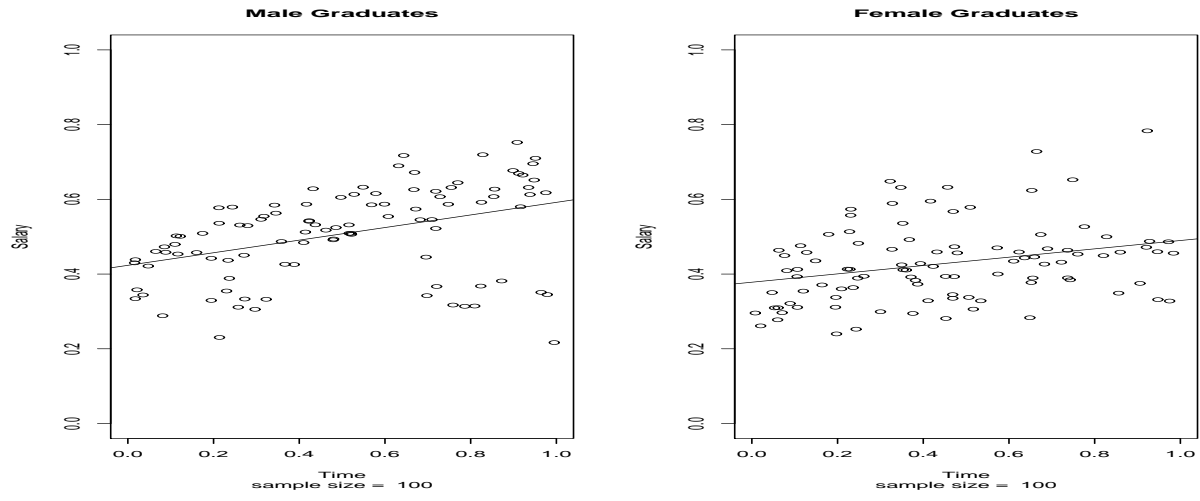


Figure 1: Data for Simpson's paradox overlaid by regression lines. Categorical variable is gender.

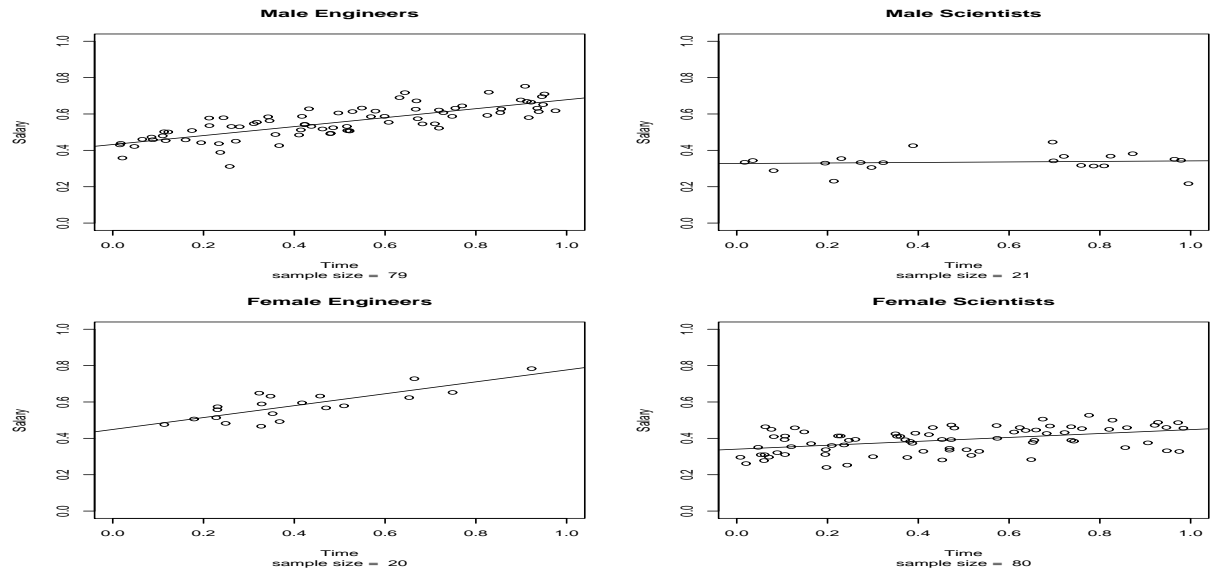


Figure 2: Data for Simpson's paradox overlaid by regression lines. Categorical variable is gender-field.

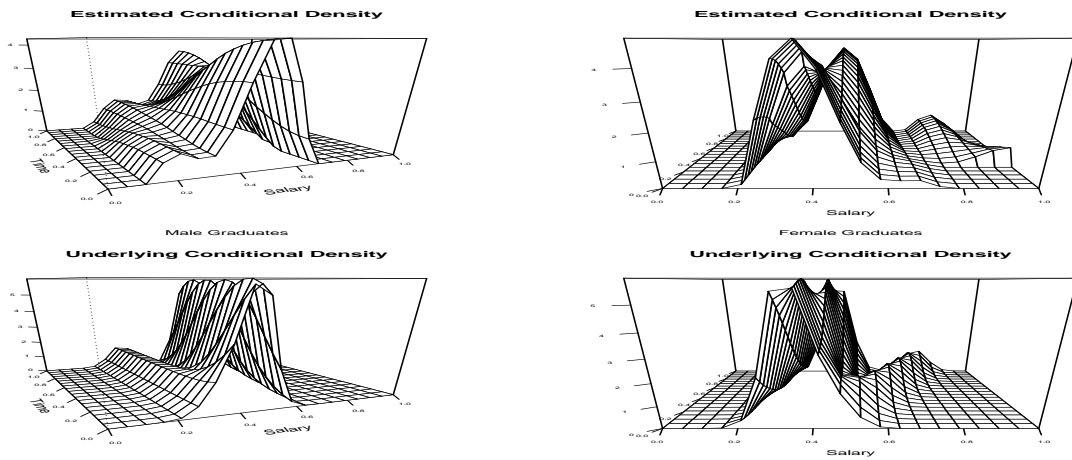


Figure 3: Estimated conditional densities of salary given time since graduation and gender.

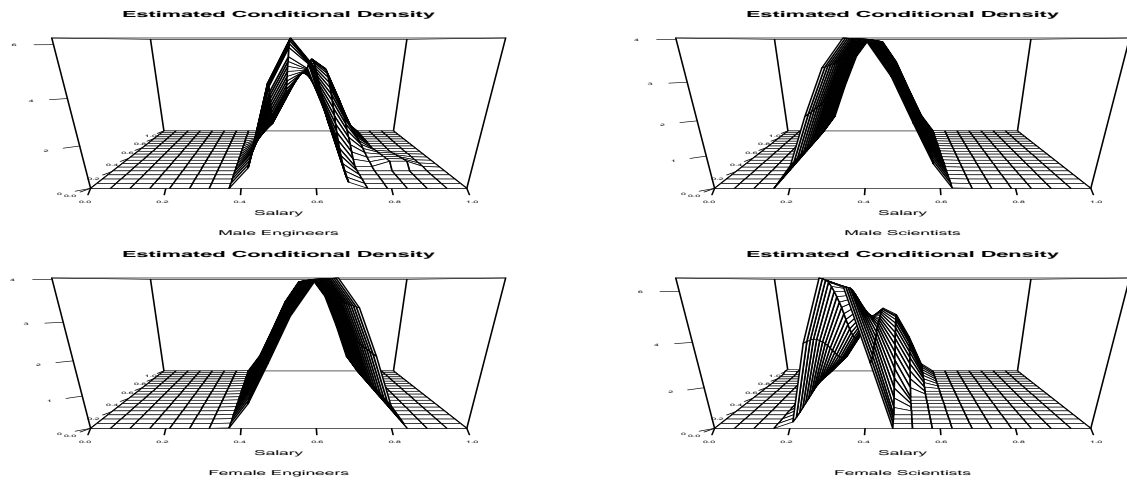


Figure 4: Estimated conditional densities of salary given time since graduation and gender-field.

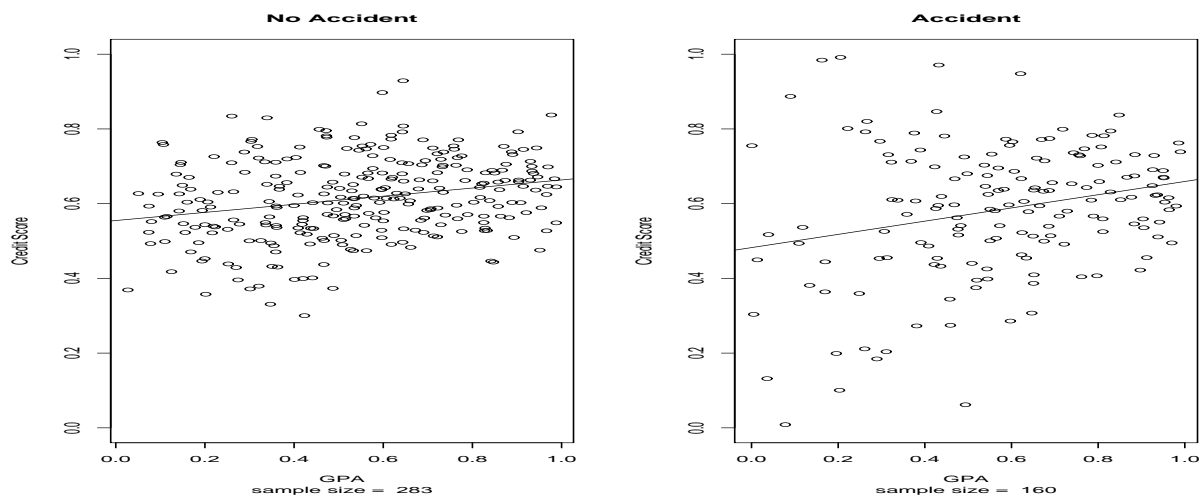


Figure 5: Scattergrams of (rescaled) credit score versus (rescaled) GPA for students with no accident and at least one accident during last 5 years. Regression lines overlay the data.

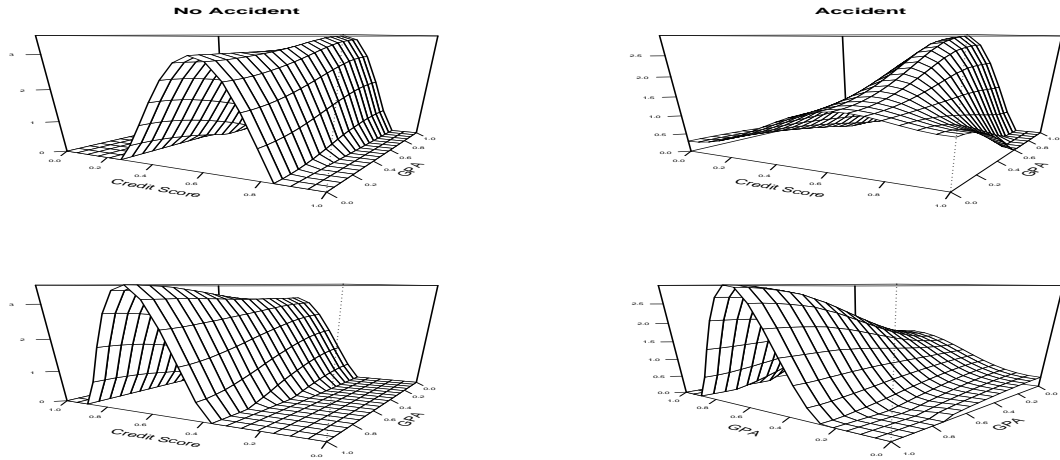


Figure 6: Estimated conditional densities of credit score given GPA and accident history. Top and bottom diagrams show the same perspective plots using different (front and back) “eye” locations.

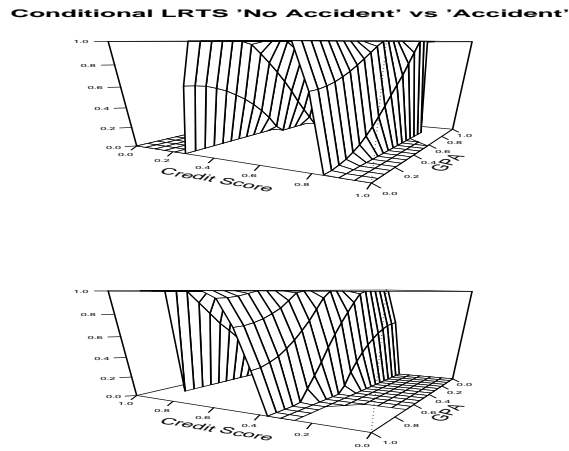


Figure 7: Conditional Likelihood Ratio Test Statistic for H_0 : No Accident versus H_a : Accident. The same perspective plot is shown using different “eye” locations.

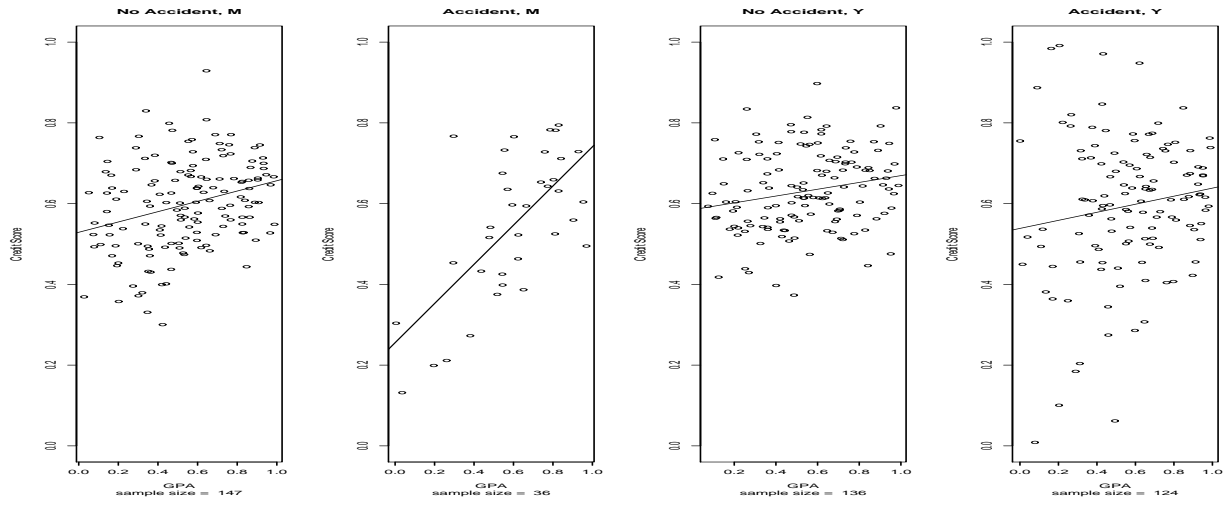


Figure 8: Credit score versus GPA for 4 categories of accident history and age. Mature (M) students are more than 22 years old and Young (Y) students are at most 22 years old.

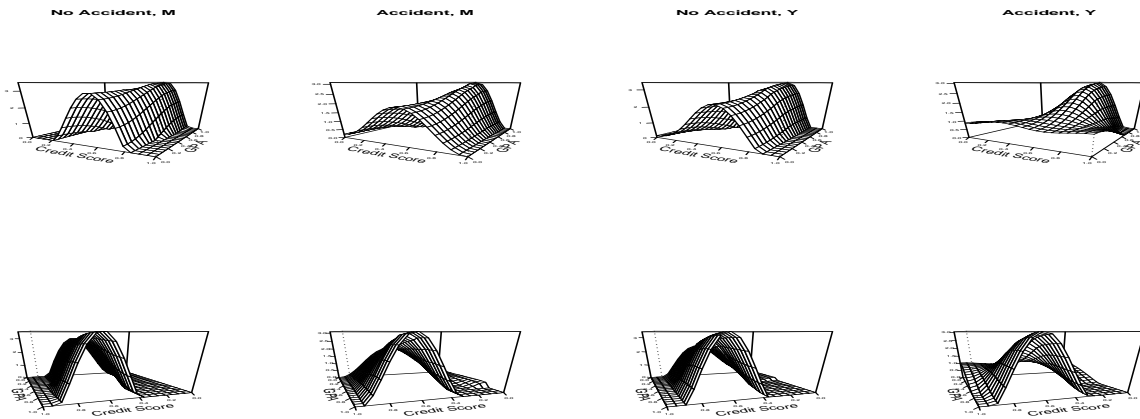
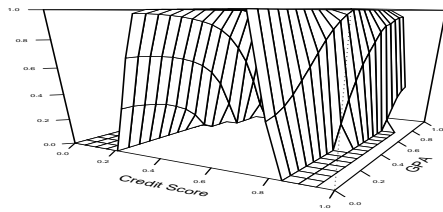


Figure 9: Conditional density estimates of credit score given GPA, accident history and age. Top and bottom diagrams show the same perspective plot using different “eye” locations.

Conditional LRTS 'No Accident,M' vs 'Accident,M'



Conditional LRTS 'No Accident,Y' vs 'Accident,Y'

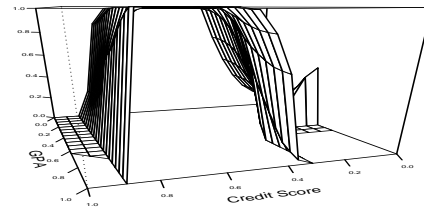
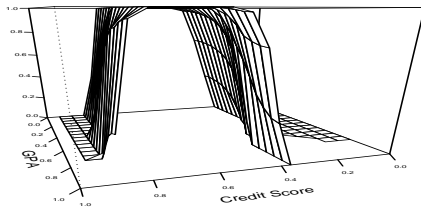
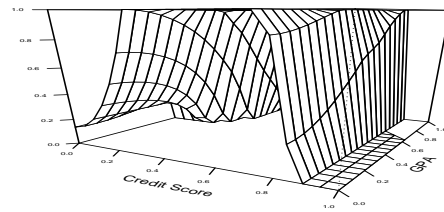


Figure 10: Conditional LRTS for testing accident history given GPA and age.