Submitted to the Annals of Statistics

EFFICIENT EMPIRICAL BAYES PREDICTION UNDER CHECK LOSS USING ASYMPTOTIC RISK ESTIMATES

By Gourab Mukherjee*, Lawrence D. Brown† and Paat Rusmevichientong*

*University of Southern California** *and University of Pennsylvania*†

We develop a novel Empirical Bayes methodology for prediction under check loss in high-dimensional Gaussian models. The check loss is a piecewise linear loss function having differential weights for measuring the amount of underestimation or overestimation. Prediction under it differs in fundamental aspects from estimation or prediction under weighted-quadratic losses. Because of the nature of this loss, our inferential target is a pre-chosen quantile of the predictive distribution rather than the mean of the predictive distribution. We develop a new method for constructing uniformly efficient asymptotic risk estimates which are then minimized to produce effective linear shrinkage predictive rules. In calculating the magnitude and direction of shrinkage, our proposed predictive rules incorporate the asymmetric nature of the loss function and are shown to be asymptotically optimal. Using numerical experiments we compare the performance of our method with traditional Empirical Bayes procedures and obtain encouraging results.

1. Introduction. We consider Empirical Bayes (EB) prediction under check loss (see Chapter 11.2.3 of Press, 2009 and Koenker and Bassett Jr, 1978) in high-dimensional Gaussian models. The check loss (sometimes also referred to as tick loss) is linear in the amount of underestimation or overestimation and the weights for these two linear segments differ. The asymmetric check loss function often arises in modern business problems as well as in medical and scientific research (Koenker, 2005). Statistical prediction analysis under asymmetric loss functions in fixed dimensional models have been considered by Zellner and Geisel (1968), Aitchison and Dunsmore (1976), Zellner (1986) and Blattberg and George (1992). Here, we consider the multivariate prediction problem under an agglomerative co-ordinatewise check loss as the dimensionality of the underlying Gaussian location model increases. In common with many other multivariate problems we find that empirical Bayes (shrinkage) can provide better performance than simple coordinate-wise rules; see James and Stein (1961), Zhang (2003), and Greenshtein and Ritov (2009) for some background. However, prediction under the loss function here differs in fundamental aspects from estimation or prediction under the

Keywords and phrases: Shrinkage estimators, Empirical Bayes prediction, Asymptotic optimality, Uniformly efficient risk estimates, Oracle inequality, Pin-ball loss, Piecewise linear loss, Hermite polynomials

weighted quadratic losses considered in most of the previous literature. This necessitates different strategies for creation of effective empirical Bayes predictors.

We begin by considering a Gaussian hierarchical Bayes structure, with unknown hyperparameters. We develop an estimate of the hyperparameters that is adapted to the shape of the concerned loss function. This estimate of the hyperparameters is then substituted in the Bayes formula to produce an EB predictive rule. This yields a co-ordinatewise prediction that we prove is overall asymptotically optimal as the dimension of the problem grows increasingly large. The hyperparameters are estimated by minimizing asymptotically efficient risk estimates. Due to the asymmetric nature of the loss function, direct construction of unbiased risk estimates which is usually done under weighted quadratic losses is difficult here. We develop a new asymptotically unbiased risk estimation method which involves an appropriate use of Hermite polynomial expansions for the relevant stochastic functions. Cai et al. (2011) used such an expansion for a different, though somewhat related, problem involving estimation of the L_1 norm of an unknown mean vector. In other respects our derivation logically resembles that of Xie, Kou and Brown (2012, 2015) who constructed empirical Bayes estimators built from an unbiased estimate of risk. However their problem involved estimation under quadratic loss, and the mathematical formulae they used provide exactly unbiased estimates of risk, and are quite different from those we develop.

The remainder of Section 1 describes our basic setup and gives formal statements of our main asymptotic results. Section 2 provides further details. It explains the general mathematical structure of our asymptotic risk estimation methodology and sketches the proof techniques used to prove the main theorems about it. Sections 4 and 5 contain further narrative to explain the proofs of the main results, but technical details are deferred to the Appendices. Section 3 reports on some simulations. These clarify the nature of our estimator and provide some confidence that it performs well even when the dimension of the model is not extremely large.

1.1. *Basic Setup.* We adopt the statistical prediction analysis framework of Aitchison and Dunsmore (1976) and Geisser (1993). We consider a one-step, n dimensional Gaussian predictive model where for each $i = 1, \ldots, n$, the observed past X_i and the unobserved future Y_i are distributed according to a normal distribution with an unknown mean θ_i; that is,

$$X_i = \theta_i + \sqrt{\nu_{p,i}} \cdot \epsilon_{1,i} \quad \text{for } i = 1, 2, \ldots, n \tag{1.1}$$

$$Y_i = \theta_i + \sqrt{\nu_{f,i}} \cdot \epsilon_{2,i} \quad \text{for } i = 1, 2, \ldots, n \ , \tag{1.2}$$

where the noise terms $\{\epsilon_{j,i} : j = 1, 2; i = 1, \ldots, n\}$ are i.i.d. from a standard normal distribution, and the past and future variances $\nu_{p,i}$, $\nu_{f,i}$ are known for all i. Note that, in multivariate notation $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_p)$ and $\boldsymbol{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_f)$ where $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_f$ are n dimensional diagonal matrices whose i^{th} entries are $\nu_{p,i}$ and $\nu_{f,i}$, respectively. If the mean θ_i were known, then the observed past X_i and future Y_i would be independent of each other.

Our objective is to compute $\hat{\boldsymbol{q}} = \{\hat{q}_i(\boldsymbol{X}) : 1 \le i \le n\}$ based on the past data $\boldsymbol{X}$ such that $\hat{\boldsymbol{q}}$ optimally predicts $\boldsymbol{Y}$. As a convention, we use bold font to denote vectors and matrices, while regular font denotes scalars. For ease of exposition, we will use $\widehat{\ }$ to denote data-dependent estimates, and we will sometimes write $\hat{\boldsymbol{q}}$ or its univariate version $\hat{q}_i$ without an explicit reference to $\boldsymbol{X}$.

When we predict the future Y_i by $\hat{q}_i$, the loss corresponding to the i^{th} coordinate is $b_i \cdot (Y_i - \hat{q}_i)^+ + h_i \cdot (\hat{q}_i - Y_i)^+$. This loss is related to the pin-ball loss function (Steinwart and Christmann, 2011), which is widely used in statistics and machine learning for estimating conditional quantiles. For each $\boldsymbol{X} = \boldsymbol{x}$, the associated predictive loss is given by

$$l_i(\theta_i, \hat{q}_i(\boldsymbol{x})) \;=\; \mathbb{E}_{Y_i \sim N(\theta_i\,,\,\nu_{f,i})}\left[b_i(Y_i - \hat{q}_i(\boldsymbol{x}))^+ + h_i(\hat{q}_i(\boldsymbol{x}) - Y_i)^+\right] \;, \tag{1.3}$$

where the expectation is taken over the distribution of the future Y_i only. We use the notation $N(\mu, \nu)$ to denote a normal random variable with mean μ and variance ν. Since Y_i is normally distributed with mean θ_i, it follows from Lemma 2.1 that

$$\begin{aligned} &l_i(\theta_i, \hat{q}_i) = \sqrt{\nu_{f,i}}\,(b_i + h_i)\, G(\,(\hat{q}_i - \theta_i)/\sqrt{\nu_{f,i}}\,,\, b_i/(b_i + h_i)\,),\ \text{where} \\ &G(w, \beta) = \phi(w) + w\Phi(w) - \beta w \quad \text{for } w \in \mathbb{R}, \beta \in [0,1]\,, \end{aligned} \tag{1.4}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal PDF and CDF, respectively. Thus, given $\boldsymbol{X}$, the cumulative loss associated with the n dimensional vector $\hat{\boldsymbol{q}}$ is

$$L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}) = \frac{1}{n}\sum_{i=1}^{n} l_i(\theta_i, \hat{q}_i)\ .$$

An Example: The Newsvendor problem. As a motivation for the check loss, consider the inventory management problem of a vendor who sells a large number of products. Consider a one-period setting, where based on the observed demand $\boldsymbol{X}$ in the previous period, the vendor must determine the stocking quantity $\hat{q}_i$ of each product in the next period. He has to balance the tradeoffs between stocking too much and incurring high inventory cost versus stocking too little and suffering lost sales. If each unit of inventory incurs a holding cost $h_i > 0$, and each unit of lost sale incurs a cost of $b_i > 0$, the vendor's cost function is given by (1.3). Usually, the lost sales cost is much higher than the inventory cost leading to a highly asymmetric loss function. This problem of determining optimal stocking levels is a classical problem in the literature on inventory management (Arrow, Harris and Marschak, 1951, Karlin and Scarf, 1958, Levi, Perakis and Uichanco, 2011, Rudin and Vahn, 2015) and is referred to as the *multivariate newsvendor problem*. In Appendix D, using a data-informed illustration on newsvendor problem we study estimation under (1.4).

Hierarchical Modeling and Predictive Risk. We want to minimize the expected loss $\mathbb{E}_{\boldsymbol{X}}\left[L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}})\right]$ over the class of estimators $\hat{\boldsymbol{q}}$ for **all** values of $\boldsymbol{\theta}$. If $\boldsymbol{\theta}$ were known, then by Lemma 2.1, the optimal prediction for each dimension i is given by $\theta_i + \sqrt{\nu_{f,i}}\,\Phi^{-1}(b_i/(b_i + h_i))$. In absence of such knowledge, we consider hierarchical modeling and the related Empirical Bayes (EB) approach (Robbins, 1964,

Zhang, 2003). This is a popular statistical method for combining information and conducting simultaneous inference on multiple parameters that are connected by the structure of the problem (Efron and Morris, 1973a,b, Good, 1980).

We consider the conjugate hierarchical model and put a prior distribution $\pi_{\eta,\tau}$ on each θ_i, under which $\theta_1, \theta_2, \ldots, \theta_n$ are i.i.d. from $N(\eta, \tau)$ distribution. Here, η and τ are the *unknown* location and scale hyperparameters, respectively. The *predictive risk* associated with our estimator $\hat{\boldsymbol{q}}$ is defined by

$$R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}) = \mathbb{E}_{\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_p)} \left[L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}) \right] \ ,$$

where the expectation is taken over $\boldsymbol{X}$. Note that the expectation over $\boldsymbol{Y}$ is already included in L via the definition of the loss ℓ_i. Because of the nature of the check loss function, our inferential target here is a pre-chosen quantile of the predictive distribution rather than the mean of the predictive distribution which is usually the object of interest in prediction under quadratic loss. By Lemma 2.2, the Bayes estimate – the unique minimizer of the integrated Bayes risk $B_n(\eta, \tau) = \int R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}) \pi_{\eta,\tau}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$ – is given for $i = 1, \ldots, n$ by

$$\hat{q}_i^{\mathsf{Bayes}}(\eta, \tau) = \alpha_i(\tau) X_i + (1 - \alpha_i(\tau))\eta + \sqrt{\nu_{f,i} + \alpha_i(\tau)\nu_{p,i}} \, \Phi^{-1}(b_i/(b_i + h_i)), \tag{1.5}$$

where, for all i, $\alpha_i(\tau) = \tau/(\tau + \nu_{p,i})$ denotes the shrinkage factor of coordinate i.

Standard parametric Empirical Bayes methods (Efron and Morris, 1973b, Lindley, 1962, Morris, 1983, Stein, 1962) suggest using the marginal distribution of $\boldsymbol{X}$ to estimate the unknown hyperparameters. In this paper, inspired by Stein's Unbiased Risk Estimation (SURE) approach of constructing shrinkage estimators (Stein, 1981), we consider an alternative estimation method. Afterwards, in Section 1.2, we show that our method outperforms standard parametric EB methods which are based on the popular maximum likelihood and method of moments approaches.

Class of Shrinkage Estimators: The Bayes estimates defined in (1.5) are based on the conjugate Gaussian prior and constitute a class of linear estimators (Johnstone, 2013). When the hyperparameters are estimated from data, they form a class of adaptive linear estimators. Note that these estimates themselves are not linear but are derived from linear estimators by the estimation of tuning parameters, which, in this case, correspond to the shrinkage factor $\alpha_i(\tau)$ and the direction of shrinkage η. Motivated by the form of the Bayes estimate in (1.5), we study the estimation problem in the following three specific classes of shrinkage estimators:

- **Shrinkage governed by Origin-centric priors:** Here, $\eta = 0$ and τ is estimated based on the past data $\boldsymbol{X}$. Shrinkage here is governed by mean-zero priors. This class of estimators is denoted by $\mathcal{S}^0 = \{\hat{\boldsymbol{q}}(\tau) \mid \tau \in [0, \infty]\}$, where for each τ, $\hat{\boldsymbol{q}}(\tau) = \{\hat{q}_i(\tau) : i = 1, \ldots, n\}$, and for all i,

  $$\hat{q}_i(\tau) = \alpha_i(\tau) X_i + \sqrt{\nu_{f,i} + \alpha_i(\tau)\nu_{p,i}} \, \Phi^{-1}\left(b_i/(b_i + h_i)\right) \ .$$

  We can generalize $\mathcal{S}^0$ by considering shrinkage based on priors with an a priori chosen location η_0. The corresponding class of shrinkage estimators $\mathcal{S}^A(\eta_0) =$

$\{\hat{\boldsymbol{q}}(\eta_0,\tau)|\tau\in[0,\infty]\}$, where η_0 is a prefixed location, consists of

$$\hat{q}_i(\eta_0,\tau)=\alpha_i(\tau)X_i+(1-\alpha_i(\tau))\eta_0+\sqrt{\nu_{f,i}+\alpha_i(\tau)\nu_{p,i}}\,\Phi^{-1}\left(b_i/(b_i+h_i)\right)\,.$$

As these estimators are location equivariant (Lehmann and Casella, 1998) the estimation problem in $\mathcal{S}^A(\eta_0)$ for any fixed η_0 reduces to an estimation problem in $\mathcal{S}^0$. Hence, we do not discuss shrinkage classes based on a priori centric priors as separate cases.

- **Shrinkage governed by Grand Mean centric priors:** In this case, $\eta=\bar{X}_n:=n^{-1}\sum_{i=1}^n X_i$, and τ is estimated based on the past data. Shrinkage here is governed by priors centering near the grand mean of the past $\boldsymbol{X}$. This class of estimators is denoted by $\mathcal{S}^G=\left\{\hat{\boldsymbol{q}}^G(\tau)\mid\tau\in[0,\infty]\right\}$, where for all $\tau\in[0,\infty]$ and $i=1,\ldots,n$,

$$\hat{q}_i^G(\tau)=\alpha_i(\tau)X_i+(1-\alpha_i(\tau))\bar{X}_n+\sqrt{\nu_{f,i}+\alpha_i(\tau)\nu_{p,i}}\,\Phi^{-1}\left(b_i/(b_i+h_i)\right)\,.$$

- **Shrinkage towards a general Data-Driven location:** In the final case, we consider the general class of shrinkage estimators where both η and τ are simultaneously estimated. We shrink towards a data-driven location while simultaneously optimizing the shrinkage factor; this class is denoted by $\mathcal{S}=\{\hat{\boldsymbol{q}}(\eta,\tau)\mid\eta\in\mathbb{R},\tau\in[0,\infty]\}$, where

$$\hat{q}_i(\eta,\tau)=\alpha_i(\tau)X_i+(1-\alpha_i(\tau))\eta+\sqrt{\nu_{f,i}+\alpha_i(\tau)\nu_{p,i}}\,\Phi^{-1}\left(b_i/(b_i+h_i)\right)\,.$$

1.2. *Main Results.* For ease of understanding, we first describe the results for the class $\mathcal{S}^0$ where the direction of shrinkage is governed by mean-zero priors so that $\eta=0$. The results for the other cases are stated afterwards; see Section 1.5. By definition, estimators in $\mathcal{S}^0$ are of the form: for $i=1,\ldots,n$,

$$\hat{q}_i(\tau)=\alpha_i(\tau)X_i+\sqrt{\nu_{f,i}+\alpha_i(\tau)\nu_{p,i}}\,\Phi^{-1}\left(b_i/(b_i+h_i)\right)\,, \tag{1.6}$$

where $\alpha_i(\tau)=\tau/(\tau+\nu_{p,i})$ is the shrinkage factor, and the tuning parameter τ varies from $[0,\infty]$. We next describe the reasonable and mild conditions that we impose on the problem structure. These assumptions mainly facilitate the rigorousness of the theoretical proofs and can be further relaxed for practical use.

Assumptions

A1. Bounded weights of the loss function. To avoid degeneracies in our loss function, which can be handled easily but require separate case by case inspections, we impose the following condition on the weights of the coordinatewise losses:

$$0<\inf_i b_i/(b_i+h_i)\le\sup_i b_i/(b_i+h_i)<1\quad\text{ and }\quad\sup_i(b_i+h_i)<\infty\,.$$

A2. Bounded parametric space. We assume that average magnitude of $\boldsymbol{\theta}$ is bounded:

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n|\theta_i|<\infty. \tag{1.7}$$

Note that, both A1 and A2 are benign structural assumptions necessary for clear statements of the proofs.

A3. Sufficient historical data. We assume the following upper bound on the ratio of the past to future variances for all the coordinates:

$$\sup_i \ \nu_{p,i}/\nu_{f,i} < 1/(4e) \ . \tag{1.8}$$

To understand the implication of this assumption, consider a multi-sample prediction problem where we observe m i.i.d. past data vectors from a n-dimensional Gaussian location model. Using sufficiency argument in the Gaussian setup, we can reduce this multi-sample past data problem to a vector problem by averaging across the m observations. The variance of the averaged model is proportional to m^{-1}, and in this case, we will have $\nu_{p,i}/\nu_{f,i} = m^{-1}$ for each i. Therefore, if we have a lot of independent historical records, the above condition will be satisfied. As such (1.8) holds if we have 11 or more independent and identically distributed past records. Conditions of this form are not new in the predictive literature, as the ratio of the past to future variability controls the role of estimation accuracy in predictive models (George, Liang and Xu, 2006, Mukherjee and Johnstone, 2015). Simulation experiments (see Section 3) suggest that the constant in (1.8) can be reduced but some condition of this form is needed. Also, to avoid degeneracies in general calculations with the loss function (which can be easily dealt by separate case analysis), we impose very mild assumptions on the variances: $\sup_i \nu_{f,i} < \infty$ and $\inf_i \nu_{p,i} > 0$.

Our Proposed Shrinkage Estimate: The predictive risk of estimators $\hat{\boldsymbol{q}}(\tau)$ of the form (1.6) is given by $R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau)) = \mathbb{E}_{\boldsymbol{\theta}}\left[L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))\right]$, where the expectation is taken over $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_p)$. Following Stein (1981), the idea of minimizing unbiased estimates of risk to obtain efficient estimates of tuning parameters has a considerable history in statistics (Efron and Morris, 1973b, George and Strawderman, 2012, Hoffmann, 2000, Stigler, 1990). However, as shown in Equation (1.4), our loss function $l(\cdot, \cdot)$ is *not* quadratic, so a direct construction of unbiased risk estimates is difficult. Instead, we approximate the risk function $\tau \mapsto R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))$ by an *Asymptotic Risk Estimator* (ARE) function $\tau \mapsto \widehat{\mathrm{ARE}}_n(\tau)$, which may be *biased*, but it approximates the true risk function *uniformly well for all* τ, particularly in large dimensions. Note that $\widehat{\mathrm{ARE}}_n(\tau)$ depends *only* on the observed $\boldsymbol{X}$ and τ and is not dependent on $\boldsymbol{\theta}$. The estimation procedure is fairly complicated and is built on a Hermite polynomial expansion of the risk. It is described in the next subsection (See (1.11)). Afterward, we show that our risk estimation method not only adapts to the data but also does a better job in adapting to the shape of the loss function when compared with the widely used Empirical Bayes MLE (EBML) or method of moments (EBMM) estimates. The main results of this paper are built on the following theorem.

THEOREM 1.1 (Uniform Point-wise Approximation of the Risk). *Under Assumptions A1 and A3, for all* $\boldsymbol{\theta}$ *satisfying Assumption A2 and for all estimates* $\hat{\boldsymbol{q}}(\tau) \in$

$\mathcal{S}^0$, we have

$$\lim_{n\to\infty} a_n^8 \Big\{ \sup_{\tau\in[0,\infty]} \mathbb{E}\big(\widehat{\mathrm{ARE}}_n(\tau) - R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))\big)^2 \Big\} \;=\; 0 \;,\; \textit{where } a_n = \log\log n$$

and the expectation is taken over the random variable $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \Sigma_p)$.

The above theorem shows that our proposed ARE method approximate the true risk in terms of mean square error uniformly well at each hyperparameter value. Next, we have a useful property of our class of predictors $\mathcal{S}^0$ which we will use along with Theorem 1.2. It shows that for each member of $\mathcal{S}^0$ the loss function $L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))$ uniformly concentrates around its expected value, which is the risk $R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))$.

THEOREM 1.2 (Uniform Concentration of the Loss around the Risk). *Under Assumption A1, for any $\boldsymbol{\theta}$ obeying Assumption A2,*

$$\lim_{n\to\infty} \mathbb{E}\Big[\sup_{\tau\in[0,\infty]} \big| R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau)) - L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau)) \big| \Big] = 0 \;.$$

The above two theorems have different flavors; Theorem 1.2 displays uniform L_1 convergence where as Theorem 1.1 shows convergence of the expected squared deviation at the rate of a_n^8 uniformly over all possible τ values. Proving the uniform L_1 convergence version of Theorem 1.1 as is usually done in establishing optimality results for estimation under quadratic loss, is difficult here due to the complicated nature of the ARE estimator. Also, the rate of convergence a_n^8 (which is used to tackle the discretization step mentioned afterwards) is not optimal and can be made better. However, it is enough for proving the optimality of our proposed method which is our main interest.

Combining the above two theorems, we see the average distance between $\widehat{\mathrm{ARE}}$ and the actual loss is asymptotically uniformly negligible and so, we expect that minimizing $\widehat{\mathrm{ARE}}$ would lead to an estimate with competitive performance. We propose an estimate of the tuning parameter τ for the class of shrinkage estimates $\mathcal{S}^0$ as follows:

(ARE Estimate) $$\hat{\tau}_n^{\mathrm{ARE}} = \operatorname*{arg\,min}_{\tau\in\Lambda_n} \widehat{\mathrm{ARE}}_n(\tau) \;.$$

where the minimization is done over a discrete sub-set Λ_n of $[0,\infty]$. Ideally, we would have liked to optimize the criterion over the entire domain $[0,\infty]$ of τ. The discretization is done for computational reasons as we minimize $\widehat{\mathrm{ARE}}$ by exhaustively evaluating it across the discrete set Λ_n which only depends on n and is independent of $\boldsymbol{x}$. Details about the construction of Λ_n is provided in Section A of the Appendix. Λ_n contains countably infinite points as $n \to \infty$. We subsequently show that the precision of our estimates is not hampered by such discretization of

the domain. To facilitate our discussion of the risk properties of our ARE Estimate, we next introduce the oracle loss (OR) hyperparameter

$$\tau_n^{\mathsf{OR}} = \underset{\tau\in[0,\infty]}{\arg\min}\, L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau)) \ .$$

Note that τ_n^{OR} depends not only on $\boldsymbol{x}$ but also on the unknown $\boldsymbol{\theta}$. Therefore, it is not an estimator. Rather, it serves as the theoretical benchmark of estimation accuracy because no estimator in $\mathcal{S}^0$ can have smaller risk than $\hat{\boldsymbol{q}}\left(\tau_n^{\mathsf{OR}}\right)$. Unlike the ARE estimate, τ_n^{OR} involves minimizing the true loss over the entire domain of τ. Note that $\hat{\boldsymbol{q}}^{\mathsf{Bayes}} \in \mathcal{S}_0$, and thus, even if the correct hyperparameter τ were known, the estimator $\hat{\boldsymbol{q}}\left(\tau_n^{\mathsf{OR}}\right)$ is as good as the Bayes estimator. The following theorem shows that our proposed estimator is asymptotically nearly as good as the oracle loss estimator.

THEOREM 1.3 (Oracle Optimality in Predictive Loss). *Under Assumptions A1 and A3, for all $\boldsymbol{\theta}$ satisfying Assumption A2 and for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{ L_n\left(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\tau}_n^{\mathrm{ARE}})\right) \ \geq \ L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau_n^{\mathsf{OR}})) \ + \ \epsilon \right\} = 0 \ .$$

The above theorem shows that the loss of our proposed estimator converges in probability to the optimum oracle value $L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau_n^{\mathsf{OR}}))$. We also show that, under the same conditions, it is asymptotically as good as τ_n^{OR} in terms of the risk (expected loss).

THEOREM 1.4 (Oracle Optimality in Predictive Risk). *Under Assumptions A1 and A3 and for all $\boldsymbol{\theta}$ satisfying Assumption A2,*

$$\lim_{n\to\infty} R_n\left(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\tau}_n^{\mathrm{ARE}})\right) - \mathbb{E}\left[L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau_n^{\mathsf{OR}})) \right] = 0 \ .$$

We extend the implications of the preceding theorems to show that our proposed estimator is as good as any other estimator in $\mathcal{S}^0$ in terms of both the loss and risk.

COROLLARY 1.1. *Under Assumptions A1 and A3, for all $\boldsymbol{\theta}$ satisfying Assumption A2, for any $\epsilon > 0$, and any estimator $\hat{\tau}_n \geq 0$,*

I. $\lim_{n\to\infty} \mathbb{P}\left\{ L_n\left(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\tau}_n^{\mathrm{ARE}})\right) \ \geq \ L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\tau}_n)) + \epsilon \right\} = 0$

II. $\lim_{n\to\infty} R_n\left(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\tau}_n^{\mathrm{ARE}})\right) - R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\tau}_n)) \ \leq \ 0.$

Next, we present two very popular, standard EB approaches for choosing estimators in $\mathcal{S}^0$. The Empirical Bayes ML (EBML) estimator $\hat{\boldsymbol{q}}(\hat{\tau}^{\mathrm{ML}})$ is built by maximizing the marginal likelihood of $\boldsymbol{X}$ while the method of moments (EBMM) estimator $\hat{\boldsymbol{q}}(\hat{\tau}^{\mathrm{MM}})$ is based on the moments of the marginal distribution of $\boldsymbol{X}$.

Following Xie, Kou and Brown (2012, Section 2) the hyperparameter estimates are given by

$$\begin{aligned} \hat{\tau}_n^{\mathrm{ML}} &= \operatorname*{arg\,min}_{\tau\in[0,\infty]} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^2}{\tau+\nu_{p,i}} + \log(\tau+\nu_{p,i})\right) \\ \hat{\tau}_n^{\mathrm{MM}} &= \max\left\{\frac{1}{n}\sum_{i=1}^{p}\left(X_i^2-\nu_{p,i}\right)\ ,\ 0\right\} \end{aligned} \quad . \tag{1.9}$$

For standard EB estimates $\hat{\boldsymbol{q}}(\hat{\tau}_n^{\mathrm{EB}})$, such as those in (1.9) the hyperparameter estimate $\hat{\tau}_n^{\mathrm{EB}}$ does not depend on the shape of the individual loss functions $\{(b_i, h_i) : 1 \le i \le n\}$. We provide a complete definition of $\widehat{\mathrm{ARE}}_n$ and $\hat{\tau}_n^{\mathrm{ARE}}$ in the next section from where it will be evident that our asymptotically optimal estimator $\hat{\tau}_n^{\mathrm{ARE}}$ depends on the ratios $\{b_i/(b_i+h_i) : 1 \le i \le n\}$ in an essential way that remains important as $n \to \infty$. Hence, even asymptotically, the ML and MM estimates do not always agree with $\hat{\tau}^{\mathrm{ARE}}$, particularly in cases when the ratios are not all the same. In the homoscedastic case it is easy to check that the loss function $L(\boldsymbol{\theta}, \hat{q}(\tau))$ has a unique minima in τ for any $\boldsymbol{\theta}$ obeying assumption A2; and so, by Theorems 1.1 and 1.3, it follows that any estimator as efficient as the OR estimator must asymptotically agree with $\hat{\tau}^{\mathrm{ARE}}$. Hence, unlike our proposed ARE based estimator, EBML and EBMM are not generally asymptotically optimal in the class of estimators $\mathcal{S}^0$. In Section 3.1, we provide an explicit numerical example to demonstrate the sub-optimal behavior of the EBML and EBMM estimators.

1.3. *Construction of Asymptotic Risk Estimates.* In this section, we describe the details for the construction of the Asymptotic Risk Estimation (ARE) function $\tau \mapsto \widehat{\mathrm{ARE}}_n(\tau)$, which is the core of our estimation methodology. The estimators in class $\mathcal{S}^0$ are coordinatewise rules, and the risk of such an estimate $\hat{q}(\tau)$ is

$$R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau)) = \frac{1}{n}\sum_{i=1}^{n} r_i(\theta_i, \hat{q}_i(\tau)) \ ,$$

where $r_i(\theta_i, \hat{q}_i(\tau))$ is the risk associated with the i^{th} coordinate. By Lemma 2.2, we have that

$$r_i(\theta_i, \hat{q}_i(\tau)) = (b_i+h_i)\sqrt{\nu_{f,i}+\nu_{p,i}\alpha_i^2(\tau)}\, G\left(c_i(\tau)+d_i(\tau)\theta_i, \tilde{b}_i\right) \ , \tag{1.10}$$

where for all i, $\alpha_i(\tau) = \tau/(\tau + \nu_{p,i})$, $\tilde{b}_i = b_i/(b_i+h_i)$, and

$$c_i(\tau) = \sqrt{\frac{1+\alpha_i(\tau)\nu_{p,i}}{1+\alpha_i(\tau)^2\nu_{p,i}}}\,\Phi^{-1}(\tilde{b}_i) \quad \text{and} \quad d_i(\tau) = -\frac{1-\alpha_i(\tau)}{\sqrt{\nu_{f,i}+\nu_{p,i}\alpha_i(\tau)^2}} \, .$$

The function $G(\cdot)$ is the same function as that associated with the predictive loss and was defined in (1.4). The dependence of $c_i(\tau)$ and $d_i(\tau)$ on τ is only through α_i. Note that, the risk $r_i(\theta_i, \hat{q}_i(\tau))$ is non-quadratic, non-symmetric and

not centered around θ_i. However, it is a C^∞ function of θ_i which we will use afterwards. We propose an estimate $\widehat{\mathrm{ARE}}_n(\tau)$ of the multivariate risk $R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))$ by using coordinate-wise estimate $\hat{T}_i(X_i, \tau)$ of $G(c_i(\tau) + d_i(\tau)\theta_i; \tilde{b}_i)$; that is,

$$\widehat{\mathrm{ARE}}_n(\tau) = \frac{1}{n}\sum_{i=1}^{n}(b_i + h_i)\sqrt{\nu_{f,i} + \nu_{p,i}\alpha_i(\tau)^2}\,\hat{T}_i(X_i, \tau)\ . \tag{1.11}$$

Defining the Coordinatewise Estimate $\hat{T}_i(X_i, \tau)$ – Heuristic Idea. Temporarily keeping the dependence on τ and i implicit, we now describe how we develop an estimate of the non-linear functional $G(c + d\,\theta, \tilde{b})$ of the unknown parameter θ.

Depending on the magnitude of $c + d\,\theta$ we use two different kinds of estimation strategy for $G(c + d\,\theta, \tilde{b})$. If $|c + d\theta|$ is not too large we approximate the functional by $G_K(c + d\,\theta, \tilde{b})$ – its K order Taylor series expansion around 0:

$$G_K(c + d\,\theta, \tilde{b}) = G(0, \tilde{b}) + G'(0, \tilde{b})(c + d\,\theta) + \phi(0)\sum_{k=0}^{K-2}\frac{(-1)^k H_k(0)}{(k+2)!}\,(c + d\,\theta)^{k+2}\,,$$

where H_k is the k^{th} order probabilists' Hermite polynomial (Thangavelu, 1993, Ch. 1.1). If $W \sim N(\mu, \nu)$ denotes a normal random variable with mean μ and variance ν, then we can construct an unbiased estimator of the truncated functional G_K by using the following property of Hermite polynomials:

$$\textit{If } W \sim N(\mu, \nu),\ \textit{then } \nu^{k/2}\,\mathbb{E}_\mu\big\{H_k(W/\sqrt{\nu})\big\} = \mu^k \textit{ for } k \geq 1. \tag{1.12}$$

Now, if $|c + d\theta|$ is large, then the truncated Taylor's expansion $G_K(\cdot)$ would not be a good approximation of $G(c + d\,\theta, \tilde{b})$. However, in that case, as shown in Lemma 2.3, we can use linear approximations with

$$G(c + d\,\theta, \tilde{b})\ \approx\ (1 - \tilde{b})(c + d\,\theta)^+ + \tilde{b}(c + d\,\theta)^-\ ,$$

and their corresponding unbiased estimates can be used. Note that for all $x \in \mathbb{R}$, $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$.

The Details. We need to combine the aforementioned estimates together in a data-driven framework. For this purpose, we use threshold estimates. We use the idea of *sample splitting*. We use the observed data to create two independent samples by adding white noise $\boldsymbol{Z} = \{Z_i : i = 1, \ldots, n\}$ and define

$$U_i = X_i + \sqrt{\nu_{p,i}}Z_i, \quad V_i = X_i - \sqrt{\nu_{p,i}}Z_i \text{ for } i = 1, \ldots, n.$$

Noting that U_i and V_i are independent, we will use V_i to determine whether or not $c_i(\tau) + d_i(\tau)\,\theta$ is large, and then use U_i to estimate $G(c_i(\tau) + d_i(\tau)\,\theta, \tilde{b})$ appropriately. For any fixed $\tau \in [0, \infty]$ and $i = 1, \ldots, n$, we transform

$$U_i(\tau) = c_i(\tau) + d_i(\tau)U_i, \quad V_i(\tau) = c_i(\tau) + d_i(\tau)V_i.$$

Note that $U_i(\tau) \sim N(c_i(\tau) + d_i(\tau)\theta_i\,,\, 2\nu_{p,i}\, d_i^2(\tau))$. By Lemma 1.12, we construct an unbiased estimate of $G_K(c_i(\tau) + d_i(\tau)\theta_i, \tilde{b}_i)$ as

$$\begin{aligned} S_i(U_i(\tau)) =& G(0,\tilde{b}_i) + G'(0,\tilde{b}_i)U_i(\tau) \\ &+ \phi(0) \sum_{k=0}^{K_n(i)-2} \frac{(-1)^k H_k(0)}{(k+2)!} \big(2\nu_{p,i} d_i^2(\tau)\big)^{(k+2)/2} H_{k+2}\bigg(\frac{U_i(\tau)}{(2\nu_{p,i} d_i^2(\tau))^{1/2}}\bigg). \end{aligned}$$

We use a further truncation on this unbiased estimate by restricting its absolute value to n. The truncated version

$$\begin{aligned} \tilde{S}_i(U_i(\tau)) &= S_i(U_i(\tau))\, I\{|S_i(U_i(\tau))| \le n\} \,+\, n\, I\{S_i(U_i(\tau)) > n\} \,-\, n\, I\{S_i(U_i(\tau)) < -n\} \\ &= \operatorname{sign}\left(S_i(U_i(\tau))\right) \min\left\{|S_i(U_i(\tau))|\ ,\ n\right\} \end{aligned}$$

is biased. But, because of its restricted growth, it is easier to control its variance, which greatly facilitates our analysis.

Threshold Estimates. For each coordinate i, we then construct the following coordinatewise threshold estimates:

$$\hat{T}_i(X_i, Z_i, \tau) = \begin{cases} -\tilde{b}_i\, U_i(\tau) & \text{if} \quad V_i(\tau) < -\lambda_n(i) \\ \tilde{S}_i(U_i(\tau)) & \text{if} \quad -\lambda_n(i) \le V_i(\tau) \le \lambda_n(i) \\ (1-\tilde{b}_i)\, U_i(\tau) & \text{if} \quad V_i(\tau) > \lambda_n(i) \end{cases} \quad \text{for } i = 1, \ldots, n$$

with the threshold parameter

$$\lambda_n(i) = \gamma(i)\sqrt{2\log n}\ , \tag{1.13}$$

where $\gamma(i)$ is any positive number less than $\big(\, 1/\sqrt{4e} - \sqrt{\nu_{p,i}/\nu_{f,i}}\,\big)$. Assumption A2 ensures the existence of $\gamma(i)$ because $\nu_{p,i}/\nu_{f,i} < 1/(4e)$ for all i.

The other tuning parameter that we have used in our construction process is the truncation parameter $K_n(i)$, which is involved in the approximation of G and is used in the estimate $\tilde{S}$. We select a choice of $K_n(i)$ that is independent of $\tau \in [0, \infty]$, and is given by

$$K_n(i) = 1 + \left\lceil e^2 \Big(\, \gamma(i) + \sqrt{2\nu_{p,i}/\nu_{f,i}}\, \Big)^2 (2\log n) \right\rceil. \tag{1.14}$$

Rao-Blackwellization. $\hat{T}_i(X_i, Z_i, \tau)$ are randomized estimators as they depend on the user-added noise $\boldsymbol{Z}$. And so, in the final step of the risk estimation procedure we apply Rao-Blackwell adjustment (Lehmann and Casella, 1998, Theorem 7.8, Page 47) to get $\hat{T}_i(X_i, \tau) = \mathbb{E}\left[\hat{T}_i(X_i, Z_i, \tau)|\boldsymbol{X}\right]$. Here, the expectation is over the distribution of $\boldsymbol{Z}$, which is independent of $\boldsymbol{X}$ and follows $N(0, I_n)$.

1.3.1. *Bias and Variance of the coordinatewise Risk Estimates.* The key result that allows us to establish Theorem 1.1 is the following proposition (for proof see

Section 2.3) on estimation of the univariate risk components $G(c_i(\tau) + d_i(\tau)\theta_i, \tilde{b}_i)$ defined in (1.10). It shows that the bias of $\hat{T}_i(X_i, Z_i, \tau)$ as an estimate of $G(c_i(\tau) + d_i(\tau)\theta_i, \tilde{b}_i)$ converges to zero as $n \to \infty$. The scaled variance of each of the univariate threshold estimates $\hat{T}_i(X_i, Z_i, \tau)$ also converges to zero.

PROPOSITION 1.1. *Under Assumptions A1 and A3, we have for all* $i = 1, \ldots, n$

I. $\lim_{n\to\infty} \sup_{\tau\in[0,\infty],\, \theta_i\in\mathbb{R}} a_n^8 \operatorname{Bias}_{\theta_i}(\hat{T}_i(X_i, Z_i, \tau)) = 0$,

II. $\lim_{n\to\infty} \sup_{\tau\in[0,\infty],\, \theta_i\in\mathbb{R}} n^{-1} a_n^8 \operatorname{Var}_{\theta_i}(\hat{T}_i(X_i, Z_i, \tau)) = 0$, *where* $a_n = \log\log n$

and the random vectors $\boldsymbol{X}$ *and* $\boldsymbol{Z}$ *are independent, with* $\boldsymbol{X}$ *following* (1.1) *and* $\boldsymbol{Z}$ *has* $N(0, I)$ *distribution.*

1.4. *Background and Previous Work.* Here, we follow the compound decision theory framework introduced in Robbins (1985). In the statistics literature, there has been substantial research on the construction of linear EB estimates in such frameworks (Morris, 1983, Zhang, 2003). Since the seminal work by James and Stein (1961), shrinkage estimators are widely used in real-world applications (Efron and Morris, 1975). Stein's shrinkage is related to hierarchical empirical Bayes methods (Stein, 1962), and several related parametric empirical Bayes estimators have been developed (Efron and Morris, 1973b). As such, Stein's Unbiased Risk Estimate (SURE) is one of the most popular methods for obtaining the estimate of tuning parameters. Donoho and Johnstone (1995) used SURE to choose the threshold parameter in their SureShrink method. However, most of these developments have been under quadratic loss or other associated loss functions (Berger, 1976, Brown, 1975, Dey and Srinivasan, 1985), which admit unbiased risk estimates. DasGupta and Sinha (1999) discussed the role of Steinien shrinkage under L_1 loss, which is related to our predictive loss only when $b = h$. If $b \neq h$, their proposed estimator do not incorporate the asymmetric nature of the loss function and are sub-optimal (See Corollary 1.1). To construct risk estimates that are adapted to the shape of the cumulative check loss functions, we develop new methods for efficiently estimating the risk functionals associated with our class of shrinkage estimators. In our construction, we concentrate on obtaining uniform convergence of the estimation error over the range of the associated hyperparameters. This enables us to efficiently fine-tune the shrinkage parameters through minimization over the class of risk estimates. Finally, in contrast to quadratic loss results (Xie, Kou and Brown, 2012, Section 3), we develop a more flexible moment-based concentration approach that translates our risk estimation efficiency into the decision theoretic optimality of the proposed shrinkage estimator.

1.5. *Further Results.* We now describe our results for efficient estimation in class $\mathcal{S}$, where we shrink towards a data-driven direction η, and the hyperparameters η and τ are simultaneously estimated. Here, we restrict the location hyperparameter η to lie in the set $\hat{M}_n = [\hat{m}_n(\alpha_1), \hat{m}_n(\alpha_2)]$ where $\hat{m}_n(\alpha) = \text{quantile}\{X_i : 1 \leq 1 \leq$

$n;\alpha\}$ is the α th quantile of the observed data and $\alpha_1 = \inf\{b_i/(b_i+h_i) : 1 \le i \le n\}$ and $\alpha_2 = \sup\{b_i/(b_i+h_i) : 1 \le i \le n\}$. By Lemma 2.1, we know that if the true distributions were known, the optimal predictor for dimension i is given by the $b_i/(b_i+h_i)$ quantile. In this context, it make sense to restrict the shrinkage location parameter η in the aforementioned range as we do not want to consider non-robust estimators that shrink toward locations that lie near undesired periphery of the data.

The predictive risk of estimators $\hat{\boldsymbol{q}}(\eta,\tau)$ of the form (1.5) is given by $R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau)) = \mathbb{E}_{\boldsymbol{\theta}}\left[L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau))\right]$. We estimate the risk function by $(\eta,\tau) \mapsto \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau)$. The estimation procedure and the detailed proof for the results in this section are presented in Section 4. We estimate the tuning parameters τ and η for the class of shrinkage estimates $\mathcal{S}$ by minimizing the $\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau)$ criterion jointly over the domain of τ and η. Computationally, it is done by minimizing over a discrete grid:

$$\left(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}}\right) = \underset{(\eta,\tau)\in(\Lambda_{n,1}\cap\hat{M}_n)\otimes\Lambda_{n,2}}{\arg\min} \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau) \ ,$$

where $\Lambda_{n,2}$ is a discrete grid spanning $[0,\infty]$ and $\Lambda_{n,1}$ is a discrete grid spanning $[-a_n, a_n]$ with $a_n = \log\log n$. Both, $\Lambda_{n,1}$ and $\Lambda_{n,2}$ do not depend on $\boldsymbol{X}$ but only on n. The minimization is conducted only over η values in $\Lambda_{n,1}$ which lie in the set $\hat{M}_n$. Details on the construction of the grid is presented in Section 4. We define the oracle loss estimator here by

$$(\eta_n^{\mathsf{DOR}},\, \tau_n^{\mathsf{DOR}}) = \underset{\tau\in[0,\infty],\ \eta\in[\hat{m}_n(\alpha_1),\hat{m}_n(\alpha_2)]}{\arg\min} L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau))$$

The following theorem shows that our risk estimates estimate the true loss uniformly well.

THEOREM 1.5. *Under Assumptions A1 and A3, for all $\boldsymbol{\theta}$ satisfying Assumption A2 and for all estimates $\hat{\boldsymbol{q}}(\eta,\tau) \in \mathcal{S}$,*

$$\lim_{n\to\infty} \sup_{\tau\in[0,\infty],\ |\eta|\le a_n} a_n^4\, \mathbb{E}\left|\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau) - L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau))\right| = 0 \qquad \textit{where } a_n = \log\log n.$$

Based on the above theorem, we derive the decision theoretic optimality of our proposed estimator. The following two theorems show that our estimator is asymptotically nearly as good as the oracle loss estimator, whereas the corollary shows that it is as good as any other estimator in $\mathcal{S}$.

THEOREM 1.6. *Under Assumptions A1 and A3, and for all $\boldsymbol{\theta}$ satisfying Assumption A2, we have, for any fixed $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left\{L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})\big) \ge L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}}, \tau_n^{\mathsf{DOR}})\big) + \epsilon\right\} = 0 \ .$$

THEOREM 1.7. *Under Assumptions A1 and A3, and for all $\boldsymbol{\theta}$ satisfying Assumption A2,*

$$\lim_{n\to\infty} R_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})\big) - \mathbb{E}\big[L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}}, \tau_n^{\mathsf{DOR}})\big)\big] = 0.$$

COROLLARY 1.2. *Under Assumptions A1 and A3, for all $\boldsymbol{\theta}$ satisfying Assumption A2 and for any estimator $\hat{\tau}_n \geq 0$ and $\hat{\eta}_n \in [\hat{m}_n(\alpha_1), \hat{m}_n(\alpha_2)]$,*

I. $\lim_{n\to\infty} P\left\{L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})\big) \geq L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n, \hat{\tau}_n)\big) + \epsilon\right\} = 0$ *for any fixed $\epsilon > 0$.*

II. $\lim_{n\to\infty} R_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})\big) - R_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n, \hat{\tau}_n)\big) \leq 0.$

The EBML estimate of the hyperparameters are given by

$$\hat{\tau}_n^{\mathrm{ML}} = \operatorname*{arg\,min}_{\tau\in[0,\infty]} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(X_i - f(\tau))^2}{\tau + \nu_{p,i}} + \log(\tau + \nu_{p,i})\right) \quad \text{and} \quad \hat{\eta}_n^{\mathrm{ML}} = f(\hat{\tau}_n^{\mathrm{ML}}) \ ,$$

where $f(\tau) = f_1(\tau)\, I\{f_1(\tau) \in [\hat{m}_n(\alpha_1), \hat{m}_n(\alpha_2)]\} + \hat{m}_n(\alpha_1)\, I\{f_1(\tau) < \hat{m}_n(\alpha_1)\} + \hat{m}_n(\alpha_2)\, I\{f_1(\tau) > \hat{m}_n(\alpha_2)\}$ and $f_1(\tau) = (\sum_{i=1}^n (\tau + \nu_{p,i})^{-1} X_i)/(\sum_{i=1}^n (\tau + \nu_{p,i})^{-1})$. The method of moments (MM) estimates are roots of the following equations:

$$\tau = \frac{1}{n-1}\left(\sum_{i=1}^{n}\big(X_i - \eta\big)^2 - (1 - 1/n)\nu_{p,i}\right)_+ \text{ and } \eta = f(\tau).$$

Unlike $(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})$, the EBML and EBMM estimates of the hyperparameters do not depend on the shape of the loss functions $\{(b_i, h_i) : 1 \leq i \leq n\}$. Thus, the EBML and EBMM estimators $\hat{\boldsymbol{q}}(\hat{\eta}^{\mathrm{ML}}, \hat{\tau}^{\mathrm{ML}})$ and $\hat{\boldsymbol{q}}(\hat{\eta}^{\mathrm{MM}}, \hat{\tau}^{\mathrm{MM}})$ do not always agree with the ARE based estimator $\hat{\boldsymbol{q}}(\hat{\eta}^{\mathrm{D}}, \hat{\tau}^{\mathrm{D}})$.

Results on Estimators in $\mathcal{S}^G$. Following (1.5), the class of estimators with shrinkage towards the Grand Mean ($\bar{\boldsymbol{X}}$) of the past observations is of the following form: for $i = 1, \ldots, n$,

$$\hat{q}_i^G(\tau) = \alpha_i(\tau) X_i + (1 - \alpha_i(\tau))\bar{\boldsymbol{X}} + (\nu_{f,i} + \alpha_i(\tau)\nu_{p,i})^{1/2}\, \Phi^{-1}(\tilde{b}_i) \ , \tag{1.15}$$

where τ varies over 0 to ∞, and $\alpha_i(\tau)$, and $\tilde{b}_i$ are defined just below Equation (1.5). For any fixed τ, unlike estimators in $\mathcal{S}$, $\hat{\boldsymbol{q}}^G(\tau)$ is no longer a coordinatewise independent rule. In Section 5, we develop an estimation strategy which estimates the loss of estimators in $\mathcal{S}^G$ uniformly well over the grid Λ_n of Section 1.2.

THEOREM 1.8. *Under Assumptions A1 and A3, for all $\boldsymbol{\theta}$ satisfying Assumption A2 and for all estimates $\hat{\boldsymbol{q}}^{\mathrm{G}}(\tau) \in \mathcal{S}^G$,*

$$\lim_{n\to\infty} \mathbb{E}\left\{\sup_{\tau\in\Lambda_n}\left|\widehat{\mathrm{ARE}}_n^{\mathrm{G}}(\tau) - L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau))\right|\right\} = 0 \ .$$

We propose an estimate $\hat{\tau}_n^{\mathrm{ARE^G}} = \arg\min_{\tau\in\Lambda_n} \widehat{\mathrm{ARE}}_n^{\mathrm{G}}(\tau)$ for the hyperparameter in this class and compare its asymptotic behavior with the oracle loss $\tau_n^{\mathsf{GOR}} =$

$\arg\min_{\tau\in[0,\infty]} L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau))$. Like the other two classes, based on Theorem 1.8, here we also derive the asymptotic optimality of our proposed estimate in terms of both the predictive risk and loss.

THEOREM 1.9. *Under Assumptions A1, A3, for all $\boldsymbol{\theta}$ satisfying Assumption A2*

(A) comparing with the oracle loss estimator, we have the following:

I. $\lim_{n\to\infty} \mathbb{P}\left\{L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\hat{\tau}_n^{\mathrm{ARE^G}})\big) \geq L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau_n^{\mathsf{GOR}})\big) + \epsilon\right\} = 0$ *for any fixed* $\epsilon > 0$.

II. $\lim_{n\to\infty} R_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\hat{\tau}_n^{\mathrm{ARE^G}})\big) - \mathbb{E}\left[L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau_n^{\mathsf{GOR}})\big)\right] = 0.$

(B) for any estimate $\hat{\tau}_n \geq 0$ of the hyperparameter, we have the following:

I. $\lim_{n\to\infty} \mathbb{P}\left\{L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\hat{\tau}_n^{\mathrm{ARE^G}})\big) \geq L_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\hat{\tau}_n)\big) + \epsilon\right\} = 0$ *for any fixed* $\epsilon > 0$.

II. $\lim_{n\to\infty} R_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\hat{\tau}_n^{\mathrm{ARE^G}})\big) - R_n\big(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\hat{\tau}_n)\big) \leq 0\,.$

1.6. *Organization of the Paper.* In Section 2, we provide a detailed explanation of the results involving the class of estimators $\mathcal{S}^0$. Treating this class as the fundamental case, through the proof of Theorem 1.1, Section 2 explains the general principle behind our asymptotic risk estimation methodology and the proof techniques used in this paper. The proofs of Theorems 1.2, 1.3 and 1.4 and Corollary 1.1 are provided in Appendix A. Section 3 discusses the performance of our prediction methodology in simulation experiments. Section 4 and its associated Appendix B provide the proofs of Theorems 1.5, 1.6 and 1.7 and Corollary 1.2, which deal with estimators in class $\mathcal{S}$. The proofs of Theorems 1.8 and 1.9 involving class $\mathcal{S}^G$ are provided in Section 5 and Appendix C. In Table 2 of the Appendix, a detailed list of the notations used in the paper is provided.

2. Proof of Theorem 1.1 and Explanation of the ARE Method. In this section, we provide a detailed explanation of the results on the estimators in $\mathcal{S}^0$. This case serves as a fundamental building block and contains all the essential ingredients involved in the general risk estimation method. In subsequent sections, the procedure is extended to $\mathcal{S}$ and $\mathcal{S}^G$. We begin by laying out the proof of Theorem 1.1. The *decision theoretic optimality results* – Theorems 1.3 and 1.4 and Corollary 1.1 – follow easily from Theorems 1.1 and 1.2; their proofs are provided in Appendix A. To prove Theorem 1.2, we use the fact that the parametric space is bounded (Assumption A2) and apply the uniform SLLN argument (Newey and McFadden, 1994, Lemma 2.4) to establish the desired concentration. The detailed proof is in the Appendix A.

2.1. *Proof of Theorem 1.1.* We will use Proposition 1.1 in the proof. Let $\widehat{\mathrm{ARE}}_n(\boldsymbol{Z}, \tau)$ denote a randomized risk estimate before the Rao-Blackwellization step in Section 1.3. For any fixed τ, $\{\hat{T}_i(X_i, Z_i, \tau) : 1 \leq i \leq n\}$ are independent of each other,

so the Bias-Variance decomposition yields

$$\begin{aligned}&\mathbb{E}\left[\left(R_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\tau))-\widehat{\mathrm{ARE}}_n(\boldsymbol{Z},\tau)\right)^2\right]\\&\le A_n\left\{\left(\frac{1}{n}\sum_{i=1}^n \mathrm{Bias}(T_i(X_i,Z_i,\tau))\right)^2+\frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(T_i(X_i,Z_i,\tau))\right\},\end{aligned}\tag{2.1}$$

where $A_n=\sup\{(b_i+h_i)^2(\nu_{f,i}+\alpha_i(\tau)\nu_{p,i}) : i=1,\ldots,n\}$ and $\alpha_i(\tau)=\tau/(\tau+\nu_{p,i})$. By Assumption A1 and A3, $\sup_n A_n<\infty$. From Proposition 1.1, both terms on the right hand side after being scaled by a_n^8 uniformly converge to 0 as $n\to\infty$. This shows that

$$\lim_{n\to\infty} a_n^8 \sup_{\tau\in[0,\infty]} \mathbb{E}\left[\left(R_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\tau))-\widehat{\mathrm{ARE}}_n(\boldsymbol{Z},\tau)\right)^2\right] \;=\; 0\;,$$

where the expectation is over the distribution of $\boldsymbol{Z}$ and $\boldsymbol{X}$. As $\widehat{\mathrm{ARE}}_n(\tau)=\mathbb{E}[\widehat{\mathrm{ARE}}_n(\boldsymbol{Z},\tau)|X]$, using Jensen's inequality for conditional expectation, we have $\mathbb{E}[(R_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\tau))-\widehat{\mathrm{ARE}}_n(\boldsymbol{Z},\tau))^2]\ge\mathbb{E}[(R_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\tau))-\widehat{\mathrm{ARE}}_n(\tau))^2]$ for any n, $\boldsymbol{\theta}$ and τ. Thus,

$$\lim_{n\to\infty} a_n^8 \sup_{\tau\in[0,\infty]} \mathbb{E}\left[\left(R_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\tau))-\widehat{\mathrm{ARE}}_n(\tau)\right)^2\right] \;=\; 0\;.$$

Thus, to complete the proof, it remains to establish Proposition 1.1, which shows that both the bias and variance converge to zero as dimension of the model increases. We undertake this in the next section. Understanding how the bias and variance is controlled will help the reader to appreciate the elaborate construction process of ARE estimates and our prescribed choices of the threshold parameter $\lambda_n(i)$ and truncation parameter $K_n(i)$.

2.2. *Proof of Proposition 1.1 Overview and Reduction to the Univariate Case.* In this section, we outline the overview of the proof techniques used to establish Proposition 1.1. It suffices to consider a generic univariate setting and consider each coordinate individually. This will simplify our analysis considerably. In addition, we will make use of the following two results about the property of the loss function G. The proof of these lemmas are given in Appendix A.

LEMMA 2.1 (Formula for the Loss Function). *If $Y\sim N(\theta,\nu)$, then*

$$\mathbb{E}_\theta\left[b(Y-q)^+ + h(q-Y)^+\right] = (b+h)\sqrt{\nu}\,G\left((q-\theta)/\sqrt{\nu}\,,\,\tilde{b}\right)\;,\tag{2.2}$$

where $\tilde{b}$=b/(b+h) and for all $w\in\mathbb{R}$ and $\beta\in[0,1]$, $G(w,\beta)=\phi(w)+w\Phi(w)-\beta w$. Also, if θ is known, the loss $l(\theta,q)$ is minimized at $q=\theta+\sqrt{\nu}\Phi^{-1}(\tilde{b})$ and the minimum value is $(b+h)\sqrt{\nu}\phi(\Phi^{-1}(\tilde{b}))$.

The next lemma gives an explicit formula for the Bayes estimator and the corresponding Bayes risk in the univariate setting.

LEMMA 2.2 (Univariate Bayes Estimator). *Consider the univariate prediction problem where the past $X \sim N(\theta, \nu_p)$, the future $Y \sim N(\theta, \nu_f)$ and $\theta \sim N(\eta, \tau)$. Consider the problem of minimizing the integrated Bayes risk. Then,*

$$\min_q \int R(\theta, q)\pi_{(\eta,\tau)}(\theta \,|\, x)d\theta \;=\; (b+h)\sqrt{\nu_f + \alpha\nu_p}\;\phi(\Phi^{-1}(\tilde{b}))\;,$$

where $\tilde{b} = b/(b+h)$, $\alpha = \tau/(\tau+\nu_p)$, and $\pi_{(\eta,\tau)}(\theta \mid x)$ is the posterior density of θ given $X = x$. Also, the Bayes estimate $\hat{q}^{\mathsf{Bayes}}(\eta,\tau)$ that achieves the above minimum is given by

$$\hat{q}^{\mathsf{Bayes}}(\eta,\tau) = \alpha x + (1-\alpha)\eta + \sqrt{\nu_f + \alpha\nu_p}\,\Phi^{-1}(\tilde{b})\;.$$

Finally, the risk $r(\theta, \hat{q}^{\mathsf{Bayes}}(\eta,\tau))$ of the Bayes estimator is

$$(b+h)\sqrt{\nu_f + \alpha^2\nu_p}\,G(c_\tau + d_\tau\,(\theta-\eta), \tilde{b})\;,$$

where

$$c_\tau = \sqrt{(1+\alpha\nu_p)/(1+\alpha^2\nu_p)}\,\Phi^{-1}(\tilde{b}) \qquad \textit{and} \qquad d_\tau = -(1-\alpha)/\sqrt{\nu_f + \alpha^2\nu_p}.$$

By Lemma 2.1, note that the loss function is scalable in ν_f. Also by Lemma 2.2, we observe that the risk calculation depends only on the ratio ν_p/ν_f and scales with $b+h$. Thus, without loss of generality, henceforth we will assume that $\nu_f = 1$, $b+h = 1$ and write $\nu = \nu_p$ and $\tilde{b} = b/(b+h) = b$. As a convention, for any number $\beta \in [0,1]$, we write $\bar{\beta} = 1-\beta$.

Reparametrization and some new notations. In order to prove the desired result, we will work with generic univariate risk estimation problems where $X \sim N(\theta, \nu)$ and $Y \sim N(\theta, 1)$. Note that Assumption A3 requires that $\nu < 1/(4e)$. For ease of presentation, we restate and partially reformulate the univariate version of the methodology stated in Section 1.3. We conduct sample splitting by adding independent Gaussian noise Z:

$$U = X + \sqrt{\nu}Z, \quad V = X - \sqrt{\nu}Z.$$

Instead of $\tau \in [0,\infty]$, we reparameterize the problem using $\alpha = \tau/(\tau+\nu) \in [0,1]$. By Lemma 2.2 and the fact that $\nu_f = 1$ and $b+h = 1$, the univariate risk function (with $\eta = 0$) is given by $\alpha \mapsto G(c_\alpha + d_\alpha\theta, b)$, where $b < 1$ and

$$c_\alpha = \Phi^{-1}(b)\sqrt{(1+\alpha\nu)/(1+\alpha^2\nu)} \qquad \text{and} \qquad d_\alpha = -\bar{\alpha}/\sqrt{1+\alpha^2\nu}\;.$$

Now, consider $U_\alpha = c_\alpha + d_\alpha U, V_\alpha = c_\alpha + d_\alpha V$ and $\theta_\alpha = c_\alpha + d_\alpha\theta$. By construction $(U_\alpha, V_\alpha) \sim N(\theta_\alpha, \theta_\alpha, 2\nu d_\alpha^2, 2\nu d_\alpha^2, 0)$ and $\alpha \mapsto G(\theta_\alpha, b)$ is estimated by the ARE estimator $\alpha \mapsto \hat{T}_{\alpha,n}(X,Z)$, where

$$\hat{T}_{\alpha,n}(X,Z) = -b\,U_\alpha\,\mathbb{I}_{\{V_\alpha < -\lambda_n\}} + \tilde{S}(U_\alpha)\,\mathbb{I}_{\{|V_\alpha| \le \lambda_n\}} + \bar{b}\,U_\alpha\,\mathbb{I}_{\{V_\alpha > \lambda_n\}}\;,$$

where $\bar{b} = 1 - b$, and the threshold is given $\lambda_n = \gamma\sqrt{2\log n}$, where γ is any positive number less than $\sqrt{2\nu}\left(\left(1/\sqrt{4e\nu}\right) - 1\right) = (1/\sqrt{2e}) - \sqrt{2\nu}$, which is well-defined by Assumption A3 because $\nu < 1/(4e)$.

The estimator $\tilde{S}(U_\alpha)$ is the truncated Taylor series expansion of $G(\theta_\alpha, b)$, defined as follows. Let

$$K_n = 1 + \left\lceil e^2\left(\gamma + \sqrt{2\nu}\right)^2 (2\log n)\right\rceil \ .$$

Let $G_{K_n}(\theta_\alpha, b)$ denote the the K_n^{th} order Taylor series expansion of $G(\theta_\alpha, b)$. Let $S(U_\alpha)$ denote an unbiased estimate of $G_{K_n}(\theta_\alpha, b)$; that is,

$$\begin{aligned} S(U_\alpha) &= G(0,b) + G'(0,b)\, U_\alpha \\ &\quad + \phi(0) \sum_{l=0}^{K_n-2} \frac{(-1)^l H_l(0)}{(l+2)!} \left(\sqrt{2\nu d_\alpha^2}\right)^{l+2} H_{l+2}\left(\frac{U_\alpha}{\sqrt{2\nu d_\alpha^2}}\right) \ , \end{aligned} \tag{2.3}$$

and finally, we have $\tilde{S}(U_\alpha) = \text{sign}(S(U_\alpha))\, \min\{|S(U_\alpha)|, n\}$, which is the truncated version of $S(U_\alpha)$. This completes the definition of the estimator $\hat{T}_{\alpha,n}(X, Z)$. This reparametrization allows us to deal with the stochasticity of the problem only through the random variables $\{U_\alpha, V_\alpha : \alpha \in [0,1]\}$ and saves us the inconvenience of dealing with the varied functionals of X and Z separately.

Proof Outline. We partition the univariate parameter space into 3 cases: `Case 1:` $|\theta_\alpha| \le \lambda_n/2$, `Case 2:` $\lambda_n/2 < |\theta_\alpha| \le \left(1 + \sqrt{2\nu}/\gamma\right)\lambda_n$ and `Case 3:` $|\theta_\alpha| > \left(1 + \sqrt{2\nu}/\gamma\right)\lambda_n$. We present a heuristic argument for considering such a decomposition. The following lemma, whose proof is provided in Appendix A, establishes a bound on the bias in different regimes.

LEMMA 2.3 (Bias Bounds). *There is an absolute constant c such that for all $b \in [0,1]$ and $\alpha \in [0,1]$,*

I. $|G(y,b) - G_{K_n}(y,b)| \le c\, \dfrac{n^{-(e^2-1)(\gamma+\sqrt{2\nu})^2}}{e^4(\gamma + \sqrt{2\nu})^2}$ for all $|y| \le \left(1 + \sqrt{2\nu}/\gamma\right)\lambda_n$.

II. $|G(y,b) - \bar{b}y| \le \dfrac{e^{-y^2/2}}{y^2}$ for all $y > 0$.

III. $|G(y,b) - (-by)| \le \dfrac{e^{-y^2/2}}{y^2}$ for all $y < 0$.

Thus, we would like to use linear estimates when $|w|$ is large and $S(U_\alpha)$ otherwise. The choice of threshold λ_n is chosen such that this happens with high probability. As we have a normal model in Case 3, which includes unbounded parametric values, we will be mainly using the linear estimates of risk because when $|\theta_\alpha| \ge \left(1 + \sqrt{2\nu}/\gamma\right)\lambda_n$, the probability of selecting $\tilde{S}$ over the linear estimates is very low. Similarly, in Case 1, we will be mainly using $\tilde{S}$. Case 2 is the buffering zone where we may use either $\tilde{S}$ or the linear estimates.

We also need to control the variances of the 3 different kind of estimates used in $\hat{T}_{\alpha,n}(X,Z)$. While the variances of the linear estimators are easily controlled, we needed to pay special attention to control the variance of $S(U_\alpha)$. In the following lemma, we exhibit an upper bound on the quadratic growth of the estimator $S(U_\alpha)$. The choice of the truncation parameter K_n in $\tilde{S}(U_\alpha)$ was done is such a way that both its bias and squared growth are controlled at the desired limits.

LEMMA 2.4 (Variance Bounds). *For any $b \in [0,1]$ and $a_n = \log\log n$,*

$$\lim_{n\to\infty} \sup_{\alpha\,:\,|\theta_\alpha| \,\le\, (1+\sqrt{2\nu}/\gamma)\lambda_n} n^{-1}\, a_n^8\, \mathbb{E}_{\theta_\alpha}\left[\{S(U_\alpha)\}^2\right] = 0\ ,$$

where the expectation is over the distribution of U_α, which has $N(\theta_\alpha, 2\nu d_\alpha^2)$ distribution for all $\alpha \in [0,1]$.

Our proof also makes use of the following large deviation bounds.

LEMMA 2.5 (Large Deviation Bounds).

For `Case 1`,
$$\lim_{n\to\infty} \sup_{\alpha\,:\,|\theta_\alpha| \,\le\, \lambda_n/2} a_n^8\, \lambda_n^2 \cdot \mathbb{P}_{\theta_\alpha}\{|V_\alpha| > \lambda_n\} = 0\ .$$

For `Case 2`,

$$\lim_{n\to\infty} \sup_{\alpha\,:\,\lambda_n/2 \,<\, \theta_\alpha \,\le\, (1+\sqrt{2\nu}/\gamma)\lambda_n} a_n^8\, |\theta_\alpha| \cdot \mathbb{P}_{\theta_\alpha}\{V_\alpha < -\lambda_n\} = 0\,.$$
$$\lim_{n\to\infty} \sup_{\alpha\,:\,-(1+\sqrt{2\nu}/\gamma)\lambda_n \,\le\, \theta_\alpha \,<\, -\lambda_n/2} a_n^8\, |\theta_\alpha| \cdot \mathbb{P}_{\theta_\alpha}\{V_\alpha > \lambda_n\} = 0\,.$$

For `Case 3`,

$$\lim_{n\to\infty} \sup_{\alpha\,:\,|\theta_\alpha| \,>\, \left(1+\sqrt{2\nu}/\gamma\right)\lambda_n} n\, a_n^8 \cdot \mathbb{P}\left\{|V_\alpha| \le \lambda_n\right\} \ = 0\,.$$
$$\lim_{n\to\infty} \sup_{\alpha\,:\,|\theta_\alpha| \,>\, \left(1+\sqrt{2\nu}/\gamma\right)\lambda_n} a_n^8\, \theta_\alpha^2 \cdot \mathbb{P}\left\{|V_\alpha| \le \lambda_n\right\} \ = 0\,.$$
$$\lim_{n\to\infty} \sup_{\alpha\,:\,\theta_\alpha \,>\, (1+\sqrt{2\nu}/\gamma)\lambda_n} a_n^8\, \theta_\alpha^2 \cdot \mathbb{P}_{\theta_\alpha}\{V_\alpha < -\lambda_n\} = 0\,.$$
$$\lim_{n\to\infty} \sup_{\alpha\,:\,\theta_\alpha \,<\, -(1+\sqrt{2\nu}/\gamma)\lambda_n} a_n^8\, \theta_\alpha^2 \cdot \mathbb{P}_{\theta_\alpha}\{V_\alpha > \lambda_n\} = 0\,.$$

The proofs of the above three lemmas are presented in Appendix A.

2.3. *Detailed Proof of Proposition 1.1.*

Bounding the Bias: As $\mathbb{E}[U_\alpha] = \theta_\alpha$, by definition $|\mathrm{Bias}_{\theta_\alpha}(\hat{T}_{\alpha,n})|$ equals

$$\left|\mathbb{E}[\tilde{S}(U_\alpha)] - G(\theta_\alpha, b)\right| \cdot \mathbb{P}\left\{|V_\alpha| \le \lambda_n\right\} + |\mathbb{E}[\tilde{S}(U_\alpha)] - G(\theta_\alpha, b)| \cdot \mathbb{P}\left\{|V_\alpha| > \lambda_n\right\}\ .$$

We will now show that each of the two terms on the RHS converges uniformly in α as n increases to infinity.

First Term: Consider θ_α in `Cases 1 and 2`; that is, $|\theta_\alpha| \le \left(1+\sqrt{2\nu}/\gamma\right)\lambda_n$. Since $\mathbb{E}S(U_\alpha) = G_{K_n}(\theta_\alpha, b)$, by our construction, we have that

$$|\mathbb{E}\tilde{S}(U_\alpha) - G(\theta_\alpha, b)| \le |\mathbb{E}\tilde{S}(U_\alpha) - \mathbb{E}S(U_\alpha)| + |G_{K_n}(\theta_\alpha, b) - G(\theta_\alpha, b)|,$$

and it follows from Lemma 2.3 that $\lim_{n\to\infty} \sup_{\alpha\,:\,|\theta_\alpha| \le \left(1+\sqrt{2\nu}/\gamma\right)\lambda_n} a_n^8 \, |G_{K_n}(\theta_\alpha, b) - G(\theta_\alpha, b)| = 0$. By Markov's Inequality,

$$|\mathbb{E}\tilde{S}(U_\alpha) - \mathbb{E}S(U_\alpha)| \le \mathbb{E}\left[|S(U_\alpha)|\,\mathbb{I}_{\{|S(U_\alpha)|\ge n\}}\right] \le \mathbb{E}\left[S^2(U_\alpha)\right]/n \ ,$$

whose a_n^8 multiplied version converges to zero uniformly in α as $n \to \infty$ by Lemma 2.4.

Now, consider `Case 3`, where $|\theta_\alpha| > \left(1+\sqrt{2\nu}/\gamma\right)\lambda_n$. By definition, $|\tilde{S}(U_\alpha)| \le n$, and by Lemma E.5, $G(\theta_\alpha, b) \le \phi(0) + \max\{\bar{b}, b\}|\theta_\alpha|$. From Lemma 2.5, we have that $\lim_{n\to\infty} \sup_{\alpha\,:\,|\theta_\alpha| > \left(1+\sqrt{2\nu}/\gamma\right)\lambda_n} a_n^8 \, \max\{n, \theta_\alpha^2\} \cdot \mathbb{P}\left\{|V_\alpha| \le \lambda_n\right\} = 0$, and thus,

$$\lim_{n\to\infty} \sup_{\alpha\,:\,|\theta_\alpha| > \left(1+\sqrt{2\nu}/\gamma\right)\lambda_n} a_n^8 \left|\mathbb{E}[\tilde{S}(U_\alpha)] - G(\theta_\alpha, b)\right| \cdot \mathbb{P}\left\{|V_\alpha| \le \lambda_n\right\} = 0 \ .$$

Therefore, in all three cases, the first term of the bias multiplied by a_n^8 converges to zero.

Second Term: The second term in the bias formula is equal to

$$B_{\alpha,n} \equiv |\bar{b}\theta_\alpha - G(\theta_\alpha, b)| \cdot \mathbb{P}\left\{V_\alpha > \lambda_n\right\} + |G(\theta_\alpha, b) - (-b\theta_\alpha)| \cdot \mathbb{P}\left\{V_\alpha < -\lambda_n\right\} \ .$$

For θ_α in `Case 1` with $|\theta_\alpha| \le \lambda_n/2$, note that by Lemma E.5,

$$\max\{|\bar{b}\theta_\alpha - G(\theta_\alpha, b)| \ , \ |G(\theta_\alpha, b) - (-b\theta_\alpha)|\} \le |\theta_\alpha| + \phi(0) + |\theta_\alpha| \le \lambda_n + \phi(0) \ ,$$

and thus $B_{\alpha,n} \le (\lambda_n + \phi(0))\,\mathbb{P}\{|V_\alpha| > \lambda_n\}$. The desired result then follows from Lemma 2.5 for Case 1.

Now, consider θ_α in `Case 2`; that is, $\lambda_n/2 < |\theta_\alpha| \le (1+\sqrt{2\nu}/\gamma)\lambda_n$. We will assume that $\lambda_n/2 < \theta_\alpha \le (1+\sqrt{2\nu}/\gamma)\lambda_n$; the case $-(1+\sqrt{2\nu}/\gamma)\lambda_n < \theta_\alpha < -\lambda_n/2$ follows analogously. Since $\theta_\alpha > \lambda_n/2$, it follows from Lemma 2.3 that

$$|\bar{b}\theta_\alpha - G(\theta_\alpha, b)| \le e^{-\theta_\alpha^2/2}/\theta_\alpha^2 \le 4\, e^{-\lambda_n^2/8}/\lambda_n^2 = 4\, n^{-\gamma^2/4}/\lambda_n^2 \ .$$

Also, by Lemma E.5, $|G(\theta_\alpha, b) - (-b\theta_\alpha)| \le 2|\theta_\alpha| + \phi(0)$. Therefore,

$$B_{\alpha,n} \le 4\, c\, n^{-\gamma^2/4}/\lambda_n^2 \ + \ (2|\theta_\alpha| + \phi(0))\mathbb{P}\left\{V_\alpha < -\lambda_n\right\} \ ,$$

and the desired result then follows from Lemma 2.5 for `Case 2`.

Now, consider θ_α in `Case 3`; that is, $|\theta_\alpha| > (1+\sqrt{2\nu}/\gamma)\lambda_n$. We will assume that $\theta_\alpha > (1+\sqrt{2\nu}/\gamma)\lambda_n$; the case $\theta_\alpha < -(1+\sqrt{2\nu}/\gamma)\lambda_n$ follows analogously. As before, it follows from Lemma 2.3 that

$$|\bar{b}\theta_\alpha - G(\theta_\alpha, b)| \leq c\, e^{-(1+\sqrt{2\nu}/\gamma)^2\lambda_n^2/2} / \big((1+\sqrt{2\nu}/\gamma)^2\lambda_n^2\big) = c\, n^{-(\gamma+\sqrt{2\nu})^2} / \big((\gamma+\sqrt{2\nu})^2(2\log n)\big)\,.$$

By Lemma E.5, $|G(\theta_\alpha, b) - (-b\theta_\alpha)| \leq 2|\theta_\alpha| + \phi(0)$. Therefore,

$$B_{\alpha,n} \leq \frac{c\, n^{-(\gamma+\sqrt{2\nu})^2}}{(\gamma+\sqrt{2\nu})^2(2\log n)} + (2|\theta_\alpha| + \phi(0))\mathbb{P}\left\{V_\alpha < -\lambda_n\right\}\ ,$$

and the desired result then follows from Lemma 2.5 for `Case 3`. Note that in `Case 3`, $|\theta_\alpha| \leq \theta_\alpha^2$ for sufficiently large n. Thus, in all three cases, the first term of the bias multiplied by a_n^8 converges to zero and we have the desired result for the bias terms in proposition.

Bounding the Variance: According to the definition of $\hat{T}_{\alpha,n}$, it follows from Lemma E.10 that

$$\begin{aligned} &\mathrm{Var}_{\theta_\alpha}(\hat{T}_{\alpha,n}) \leq 4\mathrm{Var}(A^1_{\alpha,n}) + 4\mathrm{Var}(A^2_{\alpha,n}) + 4\mathrm{Var}(A^3_{\alpha,n}),\ \text{where} \\ &A^1_{\alpha,n} = \tilde{S}(U_\alpha)\mathbb{I}_{\{|V_\alpha|<\lambda_n\}},\ A^2_{\alpha,n} = -bU_\alpha\mathbb{I}_{\{V_\alpha<-\lambda_n\}},\ \text{and}\ A^3_{\alpha,n} = \bar{b}U_\alpha\mathbb{I}_{\{V_\alpha>\lambda_n\}}. \end{aligned} \tag{2.4}$$

To establish the desired result, we will show that each term on the RHS is $o(n)$ uniformly in α; that is, for $i = 1, 2, 3$, $\lim_{n\to\infty} a_n^8\, n^{-1} \sup_{\alpha\in[0,1]} \mathrm{Var}(A^i_{\alpha,n}) = 0$.

`Case 1:` $|\theta_\alpha| \leq \lambda_n/2$. Since $\tilde{S}(U_\alpha) = \mathrm{sign}(S(U_\alpha))\, \min\{|S(U_\alpha)|, n\}$, it follows from Lemma E.3 that $\mathrm{Var}\left(A^1_{\alpha,n}\right) \leq \mathbb{E}_{\theta_\alpha}\tilde{S}^2(U_\alpha) \leq \mathbb{E}_{\theta_\alpha}S^2(U_\alpha) = o(n)$, where the last equality follows from Lemma 2.4. Again, by Lemma E.3,

$$\begin{aligned} &\mathrm{Var}(A^2_{\alpha,n}) + \mathrm{Var}(A^3_{\alpha,n}) \\ &\leq b^2\mathbb{E}\left[U_\alpha^2\right]\cdot\mathbb{P}\{V_\alpha < -\lambda_n\} + \bar{b}^2\left[U_\alpha^2\right]\cdot\mathbb{P}\{V_\alpha > \lambda_n\} \leq \mathbb{E}\left[U_\alpha^2\right]\mathbb{P}\{|V_\alpha| > \lambda_n\} \\ &= \left(\theta_\alpha^2 + 2\nu d_\alpha^2\right)\mathbb{P}\{|V_\alpha| > \lambda_n\} \leq \left(\lambda_n^2/4 + 2\nu\right)\mathbb{P}\{|V_\alpha| > \lambda_n\}\ , \end{aligned}$$

where the equality follows from the definition of U_α. The desired result then follows from Lemma 2.5 for `Case 1`.

`Case 2:` $\lambda_n/2 < |\theta_\alpha| \leq (1+\sqrt{2\nu}/\gamma)\lambda_n$. Suppose that $\lambda_n/2 < \theta_\alpha \leq (1+\sqrt{2\nu}/\gamma)\lambda_n$; the proof for the case where $-(1+\sqrt{2\nu}/\gamma)\lambda_n \leq \theta_\alpha < -\lambda_n/2$ is the same. By Lemma E.3,

$$\mathrm{Var}(A^1_{\alpha,n}) \leq \mathbb{E}\tilde{S}^2(U_\alpha) \leq \mathbb{E}S^2(U_\alpha) = o(n)\ ,$$

where the equality follows from Lemma 2.4. By Lemma E.3,

$$\mathrm{Var}(A^2_{\alpha,n}) \leq b^2\,\mathbb{E}\left[U_\alpha^2\right]\mathbb{P}\{V_\alpha < -\lambda_n\} \leq \left(2\nu + \theta_\alpha^2\right)\mathbb{P}\{V_\alpha < -\lambda_n\}\ .$$

For the range of θ_α in `Case 2`, $\theta_\alpha/n \to 0$ uniformly in α, and it follows that

$$\lim_{n\to\infty} \sup_{\alpha\,:\,\lambda_n/2 < \theta_\alpha \leq (1+\sqrt{2\nu}/\gamma)\lambda_n} a_n^8\, n^{-1}\mathrm{Var}(A^2_{\alpha,n}) = 0\,,$$

where the equality follows from Lemma 2.5 for `Case 2`. Note that $\mathrm{Var}(A^3_{\alpha,n}) \leq 4\mathbb{E}[b^2U^2_\alpha]\mathbb{P}\{V_\alpha < -\lambda_n\} \leq 4\mathbb{E}[b^2U^2_\alpha] \leq 4\left(2\nu + \theta^2_\alpha\right) = o(n)$ uniformly in α.

`Case 3`: $|\theta_\alpha| > (1+\sqrt{2\nu}/\gamma)\lambda_n$. Note that

$$\mathrm{Var}_{\theta_\alpha}(A^1_{\alpha,n}) \leq \mathbb{E}[\tilde{S}^2(U_\alpha)\mathbb{I}_{\{|V_\alpha|<\lambda_n\}}] \leq n^2\mathbb{P}\{|V_\alpha| < \lambda_n\}\ ,$$

and by Lemma 2.5 for `Case 3`, $\lim_{n\to\infty}\sup_{\alpha:|\theta_\alpha|>(1+\sqrt{2\nu}/\gamma)\lambda_n}\mathrm{Var}\theta_\alpha(A^1_{\alpha,n})/n = 0$.

Note that $\mathbb{E}[U_\alpha] = \theta_\alpha$ and $\mathrm{Var}(U_\alpha) = 2\nu d^2_\alpha \leq 2\nu$. By Lemma E.3,

$$\begin{aligned}
\mathrm{Var}_{\theta_\alpha}(A^2_{\alpha,n}) &\leq \mathbb{E}[U^2_\alpha]\,\mathbb{P}\{V_\alpha < -\lambda_n\} \leq (2\nu+\theta^2_\alpha)\mathbb{P}\{V_\alpha < -\lambda_n\}\\
\mathrm{Var}_{\theta_\alpha}(A^3_{\alpha,n}) &\leq \mathrm{Var}(\bar{b}U_\alpha) + (\mathbb{E}[\bar{b}U_\alpha])^2\mathbb{P}\{V_\alpha \leq \lambda_n\} \leq 2\nu + (2\nu+\theta^2_\alpha)\mathbb{P}\{V_\alpha \leq \lambda_n\}\\
&= 2\nu + (2\nu+\theta^2_\alpha)\mathbb{P}\{|V_\alpha| \leq \lambda_n\} + (2\nu+\theta^2_\alpha)\mathbb{P}\{V_\alpha < -\lambda_n\}\ ,
\end{aligned}$$

which implies that

$$\mathrm{Var}_{\theta_\alpha}(A^2_{\alpha,n}) + \mathrm{Var}_{\theta_\alpha}(A^3_{\alpha,n}) \leq 2\nu + (2\nu+\theta^2_\alpha)\mathbb{P}\{|V_\alpha| \leq \lambda_n\} + (2\nu+\theta^2_\alpha)\mathbb{P}\{V_\alpha < -\lambda_n\}\ .$$

Note that, by Lemma 2.5 for `Case 3`, both $\sup_{\alpha:|\theta_\alpha|>\left(1+\sqrt{2\nu}/\gamma\right)\lambda_n} a^8_n\,\theta^2_\alpha\cdot\mathbb{P}\left\{|V_\alpha| \leq \lambda_n\right\}$ and $\sup_{\alpha:\theta_\alpha>\left(1+\sqrt{2\nu}/\gamma\right)\lambda_n} a^8_n\theta^2_\alpha\cdot\mathbb{P}\left\{V_\alpha < -\lambda_n\right\}$ converge to zero as n increases. Thus, we have that $\lim_{n\to\infty}\sup_{\alpha\,:\,|\theta_\alpha|\,>\,(1+\sqrt{2\nu}/\gamma)\lambda_n} a^8_n\left(\mathrm{Var}_{\theta_\alpha}(A^2_{\alpha,n}) + \mathrm{Var}_{\theta_\alpha}(A^3_{\alpha,n})\right)/n = 0$, which is the desired result.

This completes the proof of Proposition 1.1. We end this section with a remark on the choice of threshold. The proof will work similarly for $\sqrt{2\log n}$ thresholds that are scalable with $\sqrt{\nu_{p,i}}$ and $|d_\alpha|$ for $1 \leq i \leq n$, $\alpha \in [0,1]$. Our choice λ_n being uniform over $\tau \in [0,\infty]$, however, yields a comparatively cleaner proof.

3. Simulation Experiments. In this section, we study the performances of our proposed estimators through numerical experiments. In the first example, we display a case where the performance of our proposed ARE-based estimate is close to that of the oracle estimator, but the traditional EBML and EBMM estimators perform poorly. It supports the arguments (provided below Corollaries 1.1 and 1.2) that as the formulae of the ML and MM estimates of the hyper parameters do not depend on the shape of the loss functions, they can be significantly different from the ARE-based estimates and hence sub-optimal. We calculate the inefficiency of an estimate τ of the shrinkage hyperparameter of members in $\mathcal{S}^0$ by comparing it with its corresponding Oracle risk-based estimator: $\tilde{\tau}_{OR} = \arg\min_{\tau\in[0,\infty]} R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))$. We define:

$$\text{Inefficency of } \hat{\tau} = \frac{R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\tau})) - R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tilde{\tau}_{OR}))}{\max_{\tau\geq 0} R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau)) - \min_{\tau\geq 0} R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\tau))} \times 100\,\%.$$

The measures for the other classes are defined analogously. In the other two examples, we study the performance of our proposed estimators as we vary the model parameters. Throughout this section, we set $\nu_{f,i} = 1$ and $b_i + h_1 = 1$ for all $i = 1,\ldots,n$. The R codes used for these simulation experiments can be downloaded from `http://www-bcf.usc.edu/~gourab/inventory-management/`.

3.1. *Example 1.* Here, we study a simple setup in a homoskedastic model where $\nu_{p,i} = 1/3$ for all $i = 1, \ldots, n$. We consider two different choices of n (a) $n = 20$, which yields comparatively low dimensional models, and (b) $n = 100$, which is large enough to expect our high-dimensional theory results to set in. We consider only two different values for the θ_i: $1/\sqrt{3}$ and $-3\sqrt{3}$. Also, we design the setup such that b_i is related to the θ_i: when $\theta_i = 1/\sqrt{3}$, $b_i = 0.51$ and when $\theta_i = -3\sqrt{3}$, $b_i = 0.99$. For the case when $n = 20$, we consider $(\boldsymbol{\theta}, \boldsymbol{b})$ with 18 replicates of the (θ_i, b_i) pair of $(1/\sqrt{3}, 0.51)$ and 2 replicates of $(-3\sqrt{3}, 0.99)$. For $n = 100$, we have 90 replicates of the former and 10 replicates of the latter. Note that in both the cases, the mean of $\boldsymbol{\theta}$ across dimensions is 0.

In this homoskedastic setup, the MM and ML estimates of the hyperparameter are identical. In Table 1, we present their relative inefficiencies as well as that of the ARE with respect to the Oracle risk estimate. For computation of the ARE risk estimates, 5 Monte-Carlo simulations were used for the evaluation of the unconditional expectation in the Rao-Blackwellization step. In Table 1, based on 50 independent simulation experiments, we report the mean and standard deviation of the estimates as well as their inefficiency percentages. The EBML/EBMM perform very poorly in both cases. When $n = 100$, the ARE-based estimates are close to the Oracle risk-based estimates and are quite efficient. When $n = 20$, the ARE method is not as efficient as before but still performs remarkably better than the EBML/EBMM methods. The plots of the univariate risks of $\hat{q}_i(\tau)$ for the (θ_i, b_i) pairs $(1/\sqrt{3}, 0.51)$ and $(-3\sqrt{3}, 0.99)$ (as $\alpha_i = \tau/(\tau + \nu_{p,i})$ varies) are very different (see Figure 1). For the former, the oracle minimizer is at $\alpha_{OR} = 0.51$; that is, $\tau_{OR} = 0.35$. For the latter, the oracle minimizer is at $\alpha_{OR} = 1$; that is, $\tau_{OR} = \infty$. The multivariate risk plot of our setup is different than those of the two univariate risk plots but is closer to the former than to the later. ARE approximates this multivariate risk function well and does a good job in estimating the shrinkage parameter. However, the ML/MM estimate of the hyperparameter is swayed by the extremity of fewer $(\theta_i, b_i) = (-3\sqrt{3}, 0.99)$ cases and fail to properly estimate the shrinkage parameter in the combined multivariate case.

TABLE 1

Comparison of the performances of ARE-*, MM- and ML-based estimates with the Oracle risk estimator in Example 1. The mean and standard deviation (in parentheses) across* 50 *independent simulation experiments are reported.*

METHODS	$n = 20$		$n = 100$	
	Inefficiency (%)	$\hat{\tau}$	Inefficiency (%)	$\hat{\tau}$
ARE	16.78 (30.42)	1.214 (4.823)	1.15 (2.57)	0.344 (0.079)
MM/ML	48.01 (3.55)	0.037 (0.006)	47.96 (2.01)	0.037 (0.003)
ORACLE	-	0.296 (0.000)	-	0.296 (0.000)

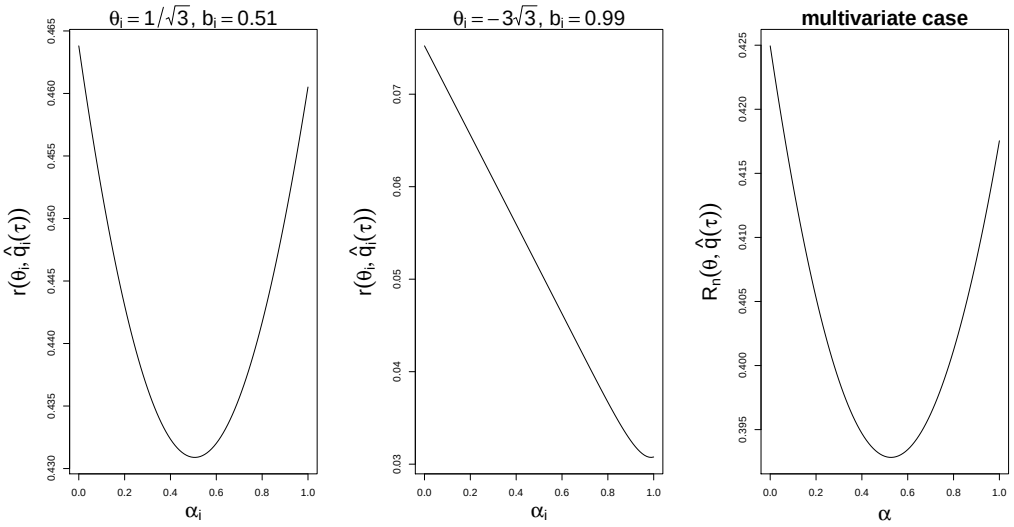

FIG 1. *From left to right we have the following: the plots of the univariate risks of* $\hat{q}_i(\tau)$ *for the* (θ_i, b_i) *pairs* $(1/\sqrt{3}, 0.51)$ *and* $(-3\sqrt{3}, 0.99)$*, respectively, as* $\alpha_i = \tau/(\tau + \nu_{p,i})$ *varies and the plot of the multivariate risk of* $\hat{\boldsymbol{q}}(\tau)$ *for the* $(\boldsymbol{\theta}, \boldsymbol{b})$ *choices described in Example 3.1.*

3.2. *Example 2.* We consider homoskedastic models with $\nu_{f,i} = 1$ and $\nu_{p,i} = \nu_p$ for all $i = 1, \ldots, n$. We vary ν_p to numerically test the performance of the ARE methodology when Assumption A3 of Section 1.2 is violated. We generate $\{\theta_i : i = 1, \ldots, n\}$ independently from $N(0, 1)$, and $\{b_i : i = 1, \ldots, n\}$ are generated uniformly from $[0.51, 0.99]$. Table 2 reports the mean and standard deviation (in brackets) of the inefficiency percentages across 20 simulation experiments from each regime. We see that the ARE methodology does not work for larger values of the ratio ν_p/ν_f and starts performing reasonably when $\nu_p/\nu_f \leq 1/3$, which is quite higher than the prescribed theoretical bound in (1.8).

TABLE 2
Inefficiency (%) of ARE *estimators in Example 2 as the ratio* ν_p/ν_f *varies.*

ν_p/ν_f	$n = 20$	$n = 100$
1/1	75.34 (28.55)	88.88 (14.70)
1/2	31.70 (20.85)	27.81 (07.95)
1/3	19.21 (14.44)	12.91 (03.63)
1/4	06.93 (03.58)	07.43 (02.07)
1/5	05.56 (03.93)	04.36 (01.38)
1/6	04.07 (03.06)	03.06 (00.97)

3.3. *Example 3.* We now study the performance of our proposed ARE^{G} methodology in 6 heteroskedastic models, which are modified predictive versions of those used in Section 7 of Xie, Kou and Brown (2012). Here, $\{b_i : i = 1, \ldots, n\}$ are generated uniformly from $[0.51, 0.99]$ and $\nu_{f,i} = 1$ for all i. Also, based on Example 2, we impose the constraint $\max\{\nu_{p,i}/\nu_{f,i} : 1 \leq i \leq n\} \leq 1/3$. Next, we outline the 6 experimental setups by describing the parameters used in the predictive model of (1.1):

Case I. $\boldsymbol{\theta}$ are i.i.d. from Uniform(0,1), and $\nu_{p,i}$ are i.i.d. from Uniform(0.1,1/3).

Case II. $\boldsymbol{\theta}$ are i.i.d. from N(0,1), and $\nu_{p,i}$ are i.i.d. from Uniform(0.1,1/3).

Case III. Here, we bring in dependence between $\nu_{p,i}$ and $\boldsymbol{\theta}$. We generate $\{\nu_{p,i} : 1 \le i \le n\}$ independently from Uniform(0.1,1/3) and $\theta_i = 5\,\nu_{p,i}$ for $i = 1, \ldots, n$.

Case IV. Instead of uniform distribution in the above case, we now generate $\{\nu_{p,i} : 1 \le i \le n\}$ independently from Inv-χ^2_{10}, which is the conjugate distribution for normal variance.

Case V. This model reflects grouping in the data. We draw the past variances independently from the 2-point distribution $2^{-1}(\delta_{0.1} + \delta_{0.5})$, and the θ_i are drawn conditioned on the past variances:

$$(\theta_i|\nu_{p,i} = 0.1) \sim N(0, 0.1) \quad \text{and} \quad (\theta_i|\nu_{p,i} = 0.5) \sim N(0, 0.5).$$

Thus, there are two groups in the data.

Case VI. In this example, we assess the sensitivity in the performance of the ARE^{G} estimators to the normality assumption by allowing $\boldsymbol{X}$ to depart from the normal model of (1.1). We generate $\{\nu_{p,i} : 1 \le i \le n\}$ independently from Uniform(0.1,1/3) and $\theta_i = 5\,\nu_{p,i}$ for $i = 1, \ldots, n$. The past observations are generated independently from

$$X_i \sim \text{Uniform}\big(\theta_i - \sqrt{3\,\nu_{p,i}},\, \theta_i + \sqrt{3\,\nu_{p,i}}\big) \text{ for } i = 1, \ldots, n.$$

Table 3 reports the mean and standard deviation (in brackets) of the inefficiency percentages of our methodology in 20 simulation experiments from each of the 6 models. The ARE^{G} estimator performs reasonably well across all 6 scenarios.

TABLE 3
Inefficiency (%) of ARE^{G} *estimators in* 6 *different heteroskedastic models of Example 3.*

	$n = 20$	$n = 100$
Case I	02.79 (02.70)	01.81 (01.83)
Case II	12.90 (21.16)	11.31 (01.73)
Case III	13.90 (19.21)	07.84 (02.08)
Case IV	08.75 (14.26)	10.47 (20.65)
Case V	03.80 (04.32)	01.52 (03.13)
Case VI	06.20 (08.45)	08.74 (03.19)

4. Explanations and Proofs for Estimators in $\boldsymbol{\mathcal{S}}$. We first describe the $\widehat{\text{ARE}}^{\text{D}}(\eta, \tau)$ risk estimation procedure. Note that, by Lemma 2.2, for any fixed $\eta \in \mathbb{R}$, the risk of estimators in $\mathcal{S}$ is related to risk of estimators in $\mathcal{S}^0$ as $R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta, \tau)) = R_n(\boldsymbol{\theta} - \eta, \hat{\boldsymbol{q}}(\tau))$. We rewrite the ARE risk estimate defined in (1.11) by explicitly denoting the dependence on $\boldsymbol{X}$ as

$$\widehat{\text{ARE}}_n(\tau, \boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}(b_i + h_i)(\nu_{f,i} + \nu_{p,i}\alpha_i^2)^{1/2}\,\hat{T}_i(X_i, \tau). \tag{4.1}$$

The $\widehat{\text{ARE}}^{\text{D}}$ risk estimate is defined as $\widehat{\text{ARE}}^{\text{D}}_n(\eta, \tau, \boldsymbol{X}) = \text{ARE}_n(\tau, \boldsymbol{X} - \eta)$. Henceforth, whenever we use the relation between $\widehat{\text{ARE}}^{\text{D}}$ and ARE, we will explicitly denote the dependence of the risk estimates on the data. Otherwise, we will stick to our earlier notation where the dependence on the data is kept implicit. We next prove Theorem 1.5.

4.1. *Proof of Theorem 1.5.* The proof follows from the following two lemmas. The first one shows that our proposed risk estimate does a good job in estimating the risk of estimators in $\mathcal{S}$. This lemma holds for all estimates $q(\eta, \tau)$ in $\mathcal{S}$ and does not need any restrictions on $\boldsymbol{\theta}$. The second lemma shows that the loss is uniformly close to the risk. It needs the restriction $|\eta| \leq a_n$ on estimates in $\mathcal{S}$ and also the assumption A2 on $\boldsymbol{\theta}$.

LEMMA 4.1. *Under Assumptions A1 and A3 with $a_n = \log\log n$, for all $\boldsymbol{\theta}$,*

$$\lim_{n\to\infty} \sup_{\tau\in[0,\infty],\, \eta\in\mathbb{R}} a_n^8\, \mathbb{E}\left[\left(\widehat{\text{ARE}}^{\text{D}}_n(\eta,\tau) - R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau))\right)^2\right] = 0\ .$$

LEMMA 4.2. *Under Assumption A1, for all $\boldsymbol{\theta}$ satisfying Assumption A2,*

$$\lim_{n\to\infty} \sup_{\tau\in[0,\infty],|\eta|\leq a_n} a_n^4\, \mathbb{E}\big|R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau)) - L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau))\big| = 0 \quad \textit{where } a_n = \log\log n.$$

The proof of Lemma 4.1 is provided in Appendix B. *For the proof of Lemma 4.2*, we show uniform convergence of the expected absolute loss over the set of location parameters $\{|\eta| \leq a_n\}$ by undertaking a moment-based approach. Here, we show that for any $\boldsymbol{\theta}$ obeying Assumption A2

$$\sup_{\tau\in[0,\infty],|\eta|\leq a_n} a_n^8 \text{Var}_{\boldsymbol{\theta}}(L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau)) \to 0 \text{ as } n \to \infty\ ,$$

from which the proof of the lemma follows easily. Now, note that, due to independence across coordinates, we have

$$\text{Var}_{\boldsymbol{\theta}}\left(L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau)\right) = n^{-2}\sum_{i=1}^n \text{Var}_{\theta_i}\left(l_i(\theta_i, \hat{q}_i(\eta,\tau))\right) \leq n^{-2}\sum_{i=1}^n \mathbb{E}_{\theta_i}\left[l_i^2(\theta_i, \hat{q}_i(\eta,\tau))\right].$$

By definition of the predictive loss, we have the following relation between the loss of estimators in $\mathcal{S}$ and $\mathcal{S}^0$: $\mathbb{E}_{\theta_i}[l_i^2(\theta_i, \hat{q}_i(\eta,\tau))] = \mathbb{E}_{\theta_i}[l_i^2(\theta_i - \alpha_i(\tau)\eta, \hat{q}_i(\tau))]$ and using the inequality in Equation (A.7) of the Appendix we see that it is dominated by $\mathcal{O}(1 + \mathbb{E}[\theta_i - \alpha_i(\tau)\eta]^2) \leq \mathcal{O}(1 + 2\,\mathbb{E}_{\theta_i}[\theta_i^2 + \eta^2])$ as $|\alpha_i(\tau)| \leq 1$ for any $\tau \in [0,\infty]$. Thus, we have

$$\text{Var}_{\boldsymbol{\theta}}\left(L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau)\right) \leq \mathcal{O}\left(n^{-2}\sum_{i=1}^n \theta_i^2 + n^{-1}a_n^2\right) \text{ for all } \tau \in [0,\infty], |\eta| \leq a_n.$$

For any $\boldsymbol{\theta}$ satisfying assumption A2, both the terms in the RHS, even after being multiplied by a_n^8 uniformly converge to 0, which completes the proof of Lemma 4.2.

We next present the proof of the decision theoretic properties of our estimators. We first define discretized set Λ_n used in the construction of the ARE estimator. $\Lambda_n = \Lambda_{n,1} \otimes \Lambda_{n,2}$ is constructed as a product grid over the space of η and τ. We consider an invertible transformation on τ and re-parametrize it by $\tilde{\tau} = \tau/(\tau+1)$. As τ varies over $[0,\infty]$, $\tilde{\tau}$ is contained in $[0,1]$. We construct an equi-spaced discretized set $\{0 = \tilde{\tau}_1 \le \tilde{\tau}_2 \le \cdots \le \tilde{\tau}_m \le 1\}$ with this transformed variable where $\tilde{\tau}_i = (i+1)\delta_{n,2}$ and $m = \lceil 1/\delta_{n,2} \rceil$ where

$$\delta_{n,2} = \left\{2\, C_1\, C_2(2\phi(0) + C_3 + \sqrt{a_n} C_4 + a_n + a_n^2)\right\}^{-1} \text{ where } a_n = \log\log n \tag{4.2}$$

and C_1, C_2 and C_3 are defined in (A.1). We retransform the aforeconstructed grid on $\tilde{\tau}$ back to get the set $\Lambda_{n,1}$ on $\tau \in [0,\infty]$. On η the grid, $\Lambda_{n,1}$ is equispaced in the interval $[-a_n, a_n]$ with the spacing equalling $\delta_{n,1} = (2C_1 a_n)^{-1}$. The cardinality of the set Λ_n is: $|\Lambda_n| = |\Lambda_{n,1}| \times |\Lambda_{n,2}| = 2a_n/\delta_{n,1} \times \delta_{n,2}^{-1} = O(a_n^4)$ as $n \to \infty$. We conduct our computations for the ARE estimate by restricting this larger grid Λ_n to the smaller set $(\Lambda_{n,1} \cap \hat{M}_n) \otimes \Lambda_{n,2}$ where the η values lie in the set $\hat{M}_n$ defined in Section 1.5.

We define the corresponding discretized version of the oracle estimator as

$$(\eta_n^\Lambda, \tau_n^\Lambda) = \underset{(\eta,\tau)\in(\Lambda_{n,1}\cap\hat{M}_n)\otimes\Lambda_{n,2}}{\arg\min} L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta,\tau)) \ .$$

The following lemma whose proof is presented in the Appendix B shows that the difference between $L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^\Lambda, \tau_n^\Lambda))$ and the original oracle loss $L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}}, \tau_n^{\mathsf{DOR}}))$ is asymptotically controlled at any prefixed level.

LEMMA 4.3. *For any fixed $\epsilon > 0$,*

I. $P\left\{L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^\Lambda, \tau_n^\Lambda)) - L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}}, \tau_n^{\mathsf{DOR}})) > \epsilon\right\} \to 0$ *as* $n \to \infty$ *and* ,

II. $\mathbb{E}|L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^\Lambda, \tau_n^\Lambda)) - L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}}, \tau_n^{\mathsf{DOR}}))| \to 0$ *as* $n \to \infty$.

We now present the proof of the decision theoretic properties of our estimators.

4.2. *Proof of Theorem 1.6.* We know that:

$$\begin{aligned}
&\mathbb{P}\left\{L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})) \ge L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}}, \tau_n^{\mathsf{DOR}})) + \epsilon\right\} \\
&\le \mathbb{P}\left\{L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})) \ge L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^\Lambda, \tau_n^\Lambda)) + \epsilon/2\right\} \\
&+ \mathbb{P}\left\{L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^\Lambda, \tau_n^\Lambda)) \ge L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}}, \tau_n^{\mathsf{DOR}})) + \epsilon/2\right\}.
\end{aligned}$$

the second term converegs to 0 by Lemma 4.3. We concentrate on the first term. Note, by construction, $\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}}) \le \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta_n^\Lambda, \tau_n^\Lambda)$. Thus, for any fixed $\epsilon > 0$, the first term is bounded above by

$$\begin{aligned}
&\mathbb{P}\left\{A_n(\boldsymbol{\theta}, \hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}}) \ge B_n(\boldsymbol{\theta}, \eta_n^\Lambda, \tau_n^\Lambda) + \epsilon/2\right\} \ , \text{ where} \\
&A_n(\boldsymbol{\theta}, \hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}}) = L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}})) - \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\hat{\eta}_n^{\mathrm{D}}, \hat{\tau}_n^{\mathrm{D}}), \text{ and} \\
&B_n(\boldsymbol{\theta}, \eta_n^\Lambda, \tau_n^\Lambda) L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}(\eta_n^\Lambda, \tau_n^\Lambda)) - \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta_n^\Lambda, \tau_n^\Lambda).
\end{aligned}$$

Now, using Markov inequality, we have

$$\mathbb{P}\left\{A_n(\boldsymbol{\theta},\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}}) \geq B_n(\boldsymbol{\theta},\eta_n^{\Lambda},\tau_n^{\Lambda}) + \epsilon/2\right\} \leq 2\epsilon^{-1}\mathbb{E}|A_n(\boldsymbol{\theta},\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}}) - B_n(\boldsymbol{\theta},\eta_n^{\Lambda},\tau_n^{\Lambda})|,$$

which, again, by the triangle inequality is less than

$$4\epsilon^{-1}\mathbb{E}\Bigg[\sup_{(\eta,\tau)\in(\Lambda_{n,1}\cap\hat{M}_n)\otimes\Lambda_{n,2}} |\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta,\tau))|\Bigg].$$

We can bound the supremum by the sum of the absolute loss over the grid and so, the above term is less than:

$$\begin{aligned}
&4\epsilon^{-1}|\Lambda_n| \sup_{(\eta,\tau)\in\Lambda_n} \mathbb{E}\Big[|\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta,\tau))|\Big] \\
&= \mathcal{O}(\epsilon^{-1}a_n^4 \sup_{\tau\in[0,\infty],|\eta|\leq a_n} \mathbb{E}\big[|\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta,\tau))|\big])
\end{aligned}$$

which, by Theorem 1.5 converges to 0 as $n\to\infty$, and we have the required result.

4.3. *Proof of Theorem 1.7.* We decompose $L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}})) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}},\tau_n^{\mathsf{DOR}}))$ into two positive components:

$$\{L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}})) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta_n^{\Lambda},\tau_n^{\Lambda}))\} + \{L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta_n^{\Lambda},\tau_n^{\Lambda})) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta_n^{\mathsf{DOR}},\tau_n^{\mathsf{DOR}}))\}.$$

The expectation of the second term converges to 0 by Lemma 4.3. For the first term we decompose the difference of the losses into 3 parts:

$$\begin{aligned}
&L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}})) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta_n^{\Lambda},\tau_n^{\Lambda}) \\
&= \left(L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}})) - \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}})\right) - \left(L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta_n^{\Lambda},\tau_n^{\Lambda})\} - \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta_n^{\Lambda},\tau_n^{\Lambda})\right) \\
&\quad + \left(\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}}) - \widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta_n^{\Lambda},\tau_n^{\Lambda})\right).
\end{aligned}$$

As the third term is less than 0, so $\mathbb{E}\left[L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\hat{\eta}_n^{\mathrm{D}},\hat{\tau}_n^{\mathrm{D}})) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta_n^{\Lambda},\tau_n^{\Lambda}))\right]$ is bounded above by $2\,\mathbb{E}\{\sup_{(\eta,\tau)\in\Lambda_n}\left|\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta,\tau))\right|\}$ which is less than

$$2\mathbb{E}\Bigg\{\sum_{(\eta,\tau)\in\Lambda_n}\left|\widehat{\mathrm{ARE}}_n^{\mathrm{D}}(\eta,\tau) - L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta,\tau))\right|\Bigg\}.$$

It converges to 0 by Theorem 1.5. Hence, the result follows.

4.4. *Proof of Corollary 1.1.* The results follow directly from Theorems 1.6 and 1.7 as $(\eta_n^{\mathsf{DOR}},\tau_n^{\mathsf{DOR}})$ minimizes the loss $L_n(\boldsymbol{\theta},\hat{\boldsymbol{q}}(\eta,\tau))$ among the class $\mathcal{S}$.

5. Explanations and Proofs for Estimators in $\mathcal{S}^G$. By (1.4), the predictive loss an estimator $\hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)$ in $\mathcal{S}^G$ is given by $L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) = \frac{1}{n}\sum_{i=1}^n l_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau))$, where

$$l_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau)) = \nu_{f,i}^{1/2}\,(b_i + h_i)\,G(\nu_{f,i}^{-1/2}(\hat{q}_i(\tau) + (1-\alpha_i)\bar{\boldsymbol{X}} - \theta_i), \tilde{b}).$$

We define a surrogate of the loss by plugging in $\bar{\boldsymbol{\theta}}$ – the mean of the unknown parameter $\boldsymbol{\theta}$ in the place of $\bar{\boldsymbol{X}}$: $\tilde{L}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) = \frac{1}{n}\sum_{i=1}^n \tilde{l}_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau))$, where

$$\tilde{l}_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau)) = \nu_{f,i}^{1/2}\,(b_i + h_i)\,G(\nu_{f,i}^{-1/2}(\hat{q}_i(\tau) + (1-\alpha_i)\bar{\boldsymbol{\theta}} - \theta_i), \tilde{b}).$$

The following lemma, whose proof is provided in Appendix C, shows the surrogate loss is uniformly close to the actual predictive loss.

LEMMA 5.1. *For any $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\hat{\boldsymbol{q}}^{\mathrm{G}}(\tau) \in \mathcal{S}^G$, we have*

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{\tau\in[0,\infty]} \left|L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - \tilde{L}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau))\right|\right] = 0\ .$$

We define the associated surrogate risk by $\tilde{r}_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau)) = \mathbb{E}_{\boldsymbol{\theta}}\tilde{l}_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau))$. From Lemma 2.2, it follows that this surrogate risk is connected with the risk function of estimators in $\mathcal{S}$ as: $\tilde{r}_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau)) = r(\theta_i - \bar{\boldsymbol{\theta}}, \hat{q}(\tau))$. Thus, the associated multivariate surrogate risk $\tilde{R}_i(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) = \sum_{i=1}^n \tilde{r}_i(\theta_i, \hat{q}_i^{\mathrm{G}}(\tau))$ equals $R_n(\theta - \bar{\boldsymbol{\theta}}, \hat{\boldsymbol{q}}(\tau))$. Also by Lemma 5.1, it follows that for any $\boldsymbol{\theta} \in \mathbb{R}^n$

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{\tau\in[0,\infty]} \left|R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - \tilde{R}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau))\right|\right] = 0\ . \tag{5.1}$$

Now we will describe our proposed $\mathrm{ARE}^{\mathrm{G}}$ estimator. Explicitly denoting the dependence of the estimators on the data, for any fixed value of $\tau \in [0,\infty]$, we define $\widehat{\mathrm{ARE}}_n^{\mathrm{G}}(\tau, \boldsymbol{X}) = \widehat{\mathrm{ARE}}_n(\tau, \boldsymbol{X} - \eta)|_{\eta=\bar{\boldsymbol{X}}}$. Note that $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$ are correlated, and $\boldsymbol{X} - \bar{\boldsymbol{X}}$ has a normal distribution with a non-diagonal covariance structure. However, we can still use the asymptotic risk estimation procedure described in Section 2 by just plugging in the value of $\bar{\boldsymbol{X}}$. We avoid the complications of incorporating the covariance structure in our calculations by cleverly using the concentration properties of $\bar{\boldsymbol{X}}$ around $\bar{\boldsymbol{\theta}}$. To explain this approach, we again define a surrogate to our $\mathrm{ARE}^{\mathrm{G}}$ estimator $\widehat{\mathrm{ARE}}_n(\tau, \boldsymbol{X} - \eta)|_{\eta=\bar{\boldsymbol{X}}} = \sum_{i=1}^n c_i\,\hat{T}_i(X_i - \eta, \tau)|_{\eta=\bar{\boldsymbol{X}}}$ by

$$\widetilde{\mathrm{ARE}}_n(\tau, \boldsymbol{X} - \bar{\boldsymbol{\theta}}) = \sum_{i=1}^n c_i\,\widetilde{\mathrm{T}}_i(X_i - \bar{\boldsymbol{\theta}}, \tau)\ ,$$

where we plugin $\bar{\boldsymbol{\theta}}$ in the place of $\bar{\boldsymbol{X}}$. Here, $c_i = (b_i + h_i)\sqrt{\nu_{f,i} + \nu_{p,i}\alpha_i(\tau)^2}$. Note that $\widetilde{\mathrm{ARE}}$ and $\widetilde{\mathrm{T}}$ have the same functional form as $\widehat{\mathrm{ARE}}$ and $\hat{T}$, respectively, but with $\bar{\boldsymbol{X}}$ replaced by $\bar{\boldsymbol{\theta}}$ and so are not estimators. We now present the proof of Theorem 1.8.

Proof of Theorem 1.8. We will prove the theorem by establishing

(a) $\lim_{n\to\infty} \mathbb{E}\{ \sup_{\tau\in\Lambda_n} \big|L_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau))\big|\} = 0$ and,

(b) $\lim_{n\to\infty} \mathbb{E}\{ \sup_{\tau\in\Lambda_n} \big|R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - \widehat{\mathrm{ARE}}^{\mathrm{G}}(\tau)\big|\} = 0.$

For the proof of (a), based on (5.1) and Lemma 5.1, it suffices to show

$$\lim_{n\to\infty} \mathbb{E}\{ \sup_{\tau\in\Lambda_n} \big|\tilde{L}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - \tilde{R}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau))\big|\} = 0\,.$$

We will prove it by showing:

$$\lim_{n\to\infty} |\Lambda_n| \sup_{\tau\in[0,\infty]} \mathbb{E}\big|\tilde{L}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - \tilde{R}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) \to 0 \text{ as } n \to \infty.$$

Recalling, $|\Lambda_n| = O(a_n)$, we show that as $n \to \infty$, $a_n^2 \mathrm{Var}_{\boldsymbol{\theta}}(\tilde{L}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)))$ converges to 0 uniformly over τ for any $\boldsymbol{\theta}$ satisfying Assumption A3 . Again, as in the proof of Lemma 4.1, we have the bound

$$\mathrm{Var}_{\boldsymbol{\theta}}(\tilde{L}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau))) \le \mathcal{O}\bigg(\frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{\theta_i}(\theta_i - \alpha_i(\tau)\bar{\boldsymbol{\theta}})^2\bigg).$$

As $|\alpha_i(\tau)| \le 1$ for all $\tau \in [0,\infty]$, the RHS above is at most $\mathcal{O}(n^{-2}\sum_{i=1}^n \theta_i^2 + \bar{\boldsymbol{\theta}}^2/n)$. Even after being scaled by a_n^2, it converges to 0 as $n \to \infty$ for any $\boldsymbol{\theta}$ satisfying Assumption A3.

Now for the proof of (b), using (5.1) as $n \to \infty$, we have

$$\begin{aligned}
&\mathbb{E}\{ \sup_{\tau\in\Lambda_n} \big|R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - \widehat{\mathrm{ARE}}^{\mathrm{G}}(\tau)\big|\} \to \mathbb{E}\{ \sup_{\tau\in\Lambda_n} \big|\tilde{R}_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathrm{G}}(\tau)) - \widehat{\mathrm{ARE}}^{\mathrm{G}}(\tau, \boldsymbol{X})\big|\} \\
&\le |\Lambda_n| \sup_{\tau\in[0,\infty]} \mathbb{E}\big|R_n(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}, \hat{\boldsymbol{q}}(\tau)) - \widehat{\mathrm{ARE}}^{\mathrm{G}}(\tau, \boldsymbol{X})\big|\,,
\end{aligned}$$

which is bounded above by the sum of $|\Lambda_n| \sup_{\tau\in[0,\infty]} \mathbb{E}_{\boldsymbol{\theta}} |\widetilde{\mathrm{ARE}}_n(\tau, \boldsymbol{X} - \bar{\boldsymbol{\theta}}) - \mathrm{ARE}^{\mathrm{G}}(\tau)|$ and $|\Lambda_n| \sup_{\tau\in[0,\infty]} \mathbb{E}_{\boldsymbol{\theta}} |R_n(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}, \hat{\boldsymbol{q}}(\tau)) - \widehat{\mathrm{ARE}}_n(\tau, \boldsymbol{X} - \bar{\boldsymbol{\theta}})|$. Again, by Lemma 4.1, the second term converges to 0 as $n \to \infty$. The first term is bounded above by

$$|\Lambda_n| \sup_{\tau\in[0,\infty]} \frac{1}{n}\sum_{i=1}^{n} c_i\, \mathbb{E}_{\boldsymbol{\theta}} \bigg| (\bar{\boldsymbol{X}} - \bar{\boldsymbol{\theta}}) \cdot \bigg[\frac{\partial}{\partial \eta} \hat{T}_i(X_i - \eta, \tau)\bigg]_{\eta=\mu_i} \bigg|\,,$$

where each $\{\mu_i : 1 \le i \le n\}$ lies between $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{X}}$. Using Cauchy-Schwarz inequality, the above term is less than

$$\lim_{n\to\infty} |\Lambda_n| \sup_{\tau\in[0,\infty]} \frac{1}{n}\sum_{i=1}^{n} c_i \bigg\{ \mathbb{E}_{\boldsymbol{\theta}}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{\theta}})^2 \cdot \mathbb{E}_{\boldsymbol{\theta}} \bigg[\frac{\partial}{\partial \eta} \hat{T}_i(X_i - \eta, \tau)\bigg]_{\eta=\mu_i}^2 \bigg\}^{1/2} = 0\,.$$

As c_i are bounded by Assumptions A1 and A3 and $|\Lambda_n| = \mathcal{O}(a_n)$, the asymptotic convergence above follows by using $\mathbb{E}_{\boldsymbol{\theta}}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{\theta}})^2 = n^{-1}$ and the following lemma, whose proof is provided in Appendix C.

LEMMA 5.2. *For any* $\boldsymbol{\theta} \in \mathbb{R}^n$ *and* μ_i *lying in between* $\bar{\boldsymbol{X}}$ *and* $\bar{\boldsymbol{\theta}}$ *for all* $i = 1, \ldots, n$

$$\lim_{n\to\infty} n^{-1} a_n^2 \left\{ \sup_{1\le i\le n} \sup_{\tau\in[0,\infty]} \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial \eta} \hat{T}_i(X_i - \eta, \tau) \right]^2_{\eta=\mu_i} \right\} = 0 \, .$$

This completes the proof of Theorem 1.8.

The proof of Theorem 1.9 follows similarly from the proofs of Theorems 1.6, 1.7 and Corollary 1.2 and is not presented here to avoid repetition.

6. Discussion. Here, we have developed an Empirical Bayes methodology for prediction in large dimensional Gaussian models. Our proposed method involves the calibration of the tuning parameters of skrinkage estimators by minimizing risk estimates that are adapted to the shape of the loss function. It produces asymptotically optimal prediction. Our risk estimation method and its proof techniques can also be used to construct optimal empirical Bayes predictive rules for general piecewise linear and related asymmetric loss functions, where we do not have any natural unbiased risk estimate. In this paper, we have worked in a high-dimensional Gaussian model. Though normality transformations exist for a wide range of high-dimensional models (Brown, 2008), future works in extending the methodology to non-Gaussian models, particularly discrete setups, would be interesting. Extending our Empirical Bayes approach from the one-period predictive setup to a multi-period setup would be another interesting future direction.

7. Supplementary Materials. Appendices A, B and C associated with Sections 2, 4 and 5 are provided in the supplementary materials (Mukherjee, Brown and Rusmevichientong, 2016). The supplement includes Appendix D which contains retail data based numerical experiments exhibiting encouraging performance of our proposed methodology when applied to the multivariate newsvendor problem. A glossary of all the notations as well as a list of all basic results used in the paper are also presented in the supplement.

References.

AITCHISON, J. and DUNSMORE, I. R. (1976). Statistical Prediction Analysis. *Bulletin of the American Mathematical Society* **82** 683–688.

ARROW, K. J., HARRIS, T. and MARSCHAK, J. (1951). Optimal inventory policy. *Econometrica* **19** 250–272.

BERGER, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *The Annals of Statistics* **4** 223–226.

BLATTBERG, R. C. and GEORGE, E. I. (1992). Estimation under profit-driven loss functions. *Journal of Business & Economic Statistics* **10** 437–444.

BROWN, L. D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters). *Journal of the American Statistical Association* **70** 417–427.

BROWN, L. D. (2008). In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Stat.* **2** 113–152.

CAI, T. T., LOW, M. G. et al. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics* **39** 1012–1041.

DasGupta, A. and Sinha, B. K. (1999). A new general interpretation of the Stein estimate and how it adapts: Applications. *Journal of Statistical Planning and Inference* **75** 247 - 268.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics* **13** 1581–1591.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90** 1200–1224.

Efron, B. and Morris, C. (1973a). Combining possibly related estimation problems. *Journal of the Royal Statistical Society. Series B (Methodological)* **35** 379–421.

Efron, B. and Morris, C. (1973b). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* **68** 117–130.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70** 311–319.

Geisser, S. (1993). *Predictive Inference. Monographs on Statistics and Applied Probability* **55**. Chapman and Hall, New York.

George, E. I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics* **34** 78–91.

George, E. I. and Strawderman, W. E. (2012). A tribute to Charles Stein. *Stat. Sci.* **27** 1–2.

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. *Trabajos de estadística y de investigación operativa* **31** 489–519.

Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality: The Third Erich L. Lehmann Symposium* **57** 266–275.

Hoffmann, K. (2000). Stein estimation—A review. *Statistical Papers* **41** 127-158.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley symposium on Mathematical Statistics and Probability* **1** 361–379.

Johnstone, I. M. (2013). Gaussian Estimation: Sequence and Wavelet Models. Version: 11 June, 2013. Available at `"http://www-stat.stanford.edu/~imj"`.

Karlin, S. and Scarf, H. (1958). Inventory models of the Arrow-Harris-Marschak type with time lag. In *Studies in the Mathematical Theory of Inventory and Production* Stanford University Press.

Koenker, R. (2005). *Quantile regression* 38. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society* 33–50.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation* **31**. Springer Science & Business Media.

Levi, R., Perakis, G. and Uichanco, J. (2011). The data-driven newsvendor problem: new bounds and insights. *Submitted to Operations Research, second revision was requested.*

Lindley, D. (1962). Discussion of the paper by Stein. *J. Roy. Statist. Soc. Ser. B* **24** 265–296.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78** 47–55.

Mukherjee, G., Brown, L. and Rusmevichientong, P. (2016). Supplement to "Asymptotic Risk Estimation and Empirical Bayes prediction under check loss".

Mukherjee, G. and Johnstone, I. M. (2015). Exact minimax estimation of the predictive density in sparse Gaussian models. *The Annals of Statistics* **43** 937–961.

Newey, W. K. and McFadden, D. (1994). Chapter 36: Large sample estimation and hypothesis testing. (R. F. Engle and D. L. McFadden, eds.). *Handbook of Econometrics* **4** 2111–2245. Elsevier, Edition 1.

Press, S. J. (2009). *Subjective and objective Bayesian statistics: principles, models, and applications* **590**. John Wiley & Sons.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics* **35** 1–20.

Robbins, H. (1985). Asymptotically subminimax solutions of compound statistical decision problems. In *Herbert Robbins Selected Papers* 7–24. Springer.

Rudin, C. and Vahn, G.-Y. (2015). The big data newsvendor: Practical insights from machine learning. *Available at SSRN 2559116*.

Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)* **24** 265–296.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9** 1135–1151.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **17** 211–225.

Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science* **5** 147–155.

Thangavelu, S. (1993). *Lectures on Hermite and Laguerre Expansions* **42**. Princeton Uni. Press.

Xie, X., Kou, S. and Brown, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association* **107** 1465–1479.

Xie, X., Kou, S. and Brown, L. D. (2015). Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *Journal of the American Statistical Association (to appear)* **45**.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association* **81** 446–451.

Zellner, A. and Geisel, M. S. (1968). Sensitivity of Control to Uncertainty and Form of the Criterion Function. *In the future of statistics, Ed Donald G. Watts* **81** 269–289.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods: invited paper. *Ann. Statist.* **31** 379–390.

Address of the First and Third authors
3670 Trousdale Parkway,
401 Bridge Hall,
University of Southern California,
Los Angeles, CA 90089-0809
E-mail: gourab@usc.edu
rusmevic@usc.edu

Address of the Second author
Department of Statistics
University of Pennsylvania
400 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104
E-mail: lbrown@wharton.upenn.edu