Example 1.5 The curse of dimensionality; continuation of Example 1.2 Suppose that $X^*$ has many continuous components and we assume that $E(\varepsilon(\mu)\varepsilon(\mu)^T | X^*)$ is unrestricted except for being a continuous function of the continuous components of $X^*$. To simplify the notation, we assume $\mu$ is one-dimensional. It is possible to use multivariate smoothing to construct a globally efficient RAL estimator $\mu_n, \text{globeff}$ of $\mu$ under the standard asymptotic theory of Bickel, Klaassen, Ritov, and Wellner (1993) by using a smooth to obtain a globally consistent estimator of $E(\varepsilon(\mu)\varepsilon(\mu)^T | X^*)$; that is, the estimator will have asymptotic variance equal to the semiparametric variance bound $I^{-1}(\mu, \eta)$ (i.e., the inverse of the variance of the efficient score) at each law $(\mu, \eta)$ allowed by the model. However, in finite samples and regardless of the choice of smoothing parameter (e.g., bandwidth), the actual coverage rate of the Wald interval $\mu_n, \text{globeff} \pm z_{\alpha/2} I^{-1/2}(\mu, \eta) / \sqrt{n}$ based on $\mu_n, \text{globeff}$ and the semiparametric variance bound $I^{-1}(\mu, \eta)$ will be considerably less than its nominal $(1 - \alpha)$ level at laws $(\mu, \eta)$ at which $E(\varepsilon(\mu)\varepsilon(\mu)^T | X^*)$ is a very wiggly function of $X^*$. Here $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard normal. This is because (i) if a large bandwidth is used, the estimate of $E(\varepsilon(\mu)\varepsilon(\mu)^T | X^*)$ will be biased so that $h_{\text{opt}}$ and its estimate will differ greatly but (ii) if a small bandwidth is used, the second-order $\text{op}(1/\sqrt{n})$ terms in the asymptotic linearity expansion $\mu_n, \text{globeff} - \mu = O(1/\sqrt{n})$, where $I(Y)$ denotes the influence curve of $\mu_n, \text{globeff}$, will be large, adding variability. Thus, standard asymptotics is a poor guide to finite sample performance in high-dimensional models when $X^*$ has many continuous components. Robins and Ritov (1997) proposed an alternative cure of dimensionality appropriate (CODA) asymptotics that serves as a much better guide.

Under CODA asymptotics, an estimator $\mu_n, \text{globeff}$ of a one-dimensional parameter $\mu$ is defined to be globally CODA-efficient if $\mu_n, \text{globeff} \pm z_{\alpha/2} I^{-1/2}(\mu, \eta) / \sqrt{n}$ (or equivalently $\mu_n, \text{globeff} \pm z_{\alpha/2} \tilde{I}^{-1/2}(\mu, \eta) / \sqrt{n}$, where $\tilde{I}$ is a uniformly consistent estimator of $I$) is an asymptotic $(1 - \alpha)$ confidence interval for $\mu$ uniformly over all laws $(\mu, \eta)$ allowed by the model. An estimator $\mu_n, \text{loc-eff}$ is locally CODA-efficient at a working submodel if (i) $\mu_n, \text{loc-eff} \pm z_{\alpha/2} I^{-1/2}(\mu, \eta) / \sqrt{n}$ is an asymptotic $(1 - \alpha)$ confidence interval for $\mu$ uniformly over all laws $(\mu, \eta)$ in the submodel and (ii) $\mu_n, \text{loc-eff} \pm z_{\alpha/2} \sigma_n$ is an asymptotic $(1 - \alpha)$ confidence interval for $\mu$ uniformly over all laws $(\mu, \eta)$, where $\sigma_n$ is the nonparametric bootstrap estimator of the standard error of $\mu_n, \text{loc-eff}$ (or any other robust estimator of its asymptotic standard error). Given (ii), condition (i) is implied by $\sqrt{n} \sigma_n$ converging to $I^{-1/2}$ uniformly over $(\mu, \eta)$ in the working submodel. These definitions extend to a vector parameter $\mu$ by requiring that they hold for each one-dimensional linear combination $\mu$ of the components. Arguments similar to those in Robins and Ritov (1997) show that in the model with $E(\varepsilon(\mu)\varepsilon(\mu)^T | X^*)$ unrestricted, except by continuity and a bound on its matrix norm, no globally efficient CODA estimators exist (owing to undercoverage under certain laws $(\mu, \eta)$ depending on the sample size $n$), but the locally efficient RAL estimator of the previous paragraph is locally CODA-efficient as well. Further, in moderate-sized samples, the nominal $1 - \alpha$ Wald interval confidence interval $\mu_n, \text{loc-eff} \pm z_{\alpha/2} \sigma_n$ for $\mu$ based on the locally efficient estimator above and its estimated variance will cover at near its nominal rate under all laws allowed by the model, with length near $2z_{\alpha/2} I^{-1/2}(\mu, \eta) / \sqrt{n}$ at laws in the working submodel. Thus, CODA asymptotics is much more reliable than standard asymptotics as a guide to finite sample performance.

Now $\mu_n, \text{globeff}$ can be made globally CODA-efficient if we impose the additional assumption that $E(\varepsilon(\mu)\varepsilon(\mu)^T | X^*)$ is locally smooth (i.e., has bounded derivatives to a sufficiently high order) in the continuous components of $X^*$. However, when $X^*$ is high-dimensional, even when local smoothness is known to be correct, the asymptotics based on the larger model that only assumes continuity of the conditional covariance provides a more relevant and appropriate guide to moderate sample performance. For example, with moderate-sized samples, for any estimator $\mu_n, \text{globeff}$, there will exist laws $(\mu, \eta)$ satisfying the local smoothness assumption such that the coverage of $\mu_n, \text{globeff} \pm z_{\alpha/2} I^{-1/2}(\mu, \eta) / \sqrt{n}$ will be considerably less than its nominal $(1 - \alpha)$. This is due to the curse of dimensionality: in high-dimensional models with moderate sample sizes, local smoothness assumptions, even when true, are not useful, since essentially no two units will have $X^*$-vectors close enough to one another to allow the "borrowing of information" necessary for smoothing. Thus, in high-dimensional models, we suggest using a CODA asymptotics that does not impose smoothness, even when smoothness is known to hold.

1.2. Tour through the General Estimation Problem

The distribution of $Y$ is indexed by the distribution $F_X$ of the full data structure $X$ and the conditional distribution $G(\cdot | X)$ of the censoring variable $C$, given $X$. Because, for a given $X$, the outcome of $C$ determines what we observe about $X$, we refer to the conditional distribution $G(\cdot | X)$ as the censoring or coarsening mechanism. If the censoring variable $C$ is allowed to depend on unobserved components of $X$, then $\mu$ is typically not identifiable from the distribution of $Y$ without additional strong untenable assumptions. When the censoring distribution only depends on the observed components of $X$, we say that the censoring mechanism satisfies coarsening at random (CAR).

In this book we will assume that the censoring mechanism $G$ satisfies CAR. Formally, CAR is a restriction on the conditional distribution $G_{Y|X}$ of $Y$, given $X$ (which implies that it is also a restriction on $G$). If $Y$ includes the censoring variable $C$ itself as a component, then the conditional distribution $G_{Y|X}$ of $Y$, given $X$, can be replaced by $G$ itself in the definition of