Example 1.5 The curse of dimensionality; continuation of Example 1.2 Suppose that X^* has many continuous components and we assume that $E(\epsilon(\mu)\epsilon(\mu)^\top \mid X^*)$ is unrestricted except for being a continuous function of the continuous components of X^*. To simplify the notation, we assume μ is one-dimensional. It is possible to use multivariate smoothing to construct a globally efficient RAL estimator $\mu_{n,globeff}$ of μ under the standard asymptotic theory of Bickel, Klaassen, Ritov and Wellner (1993) by using a smooth to obtain a globally consistent estimator of $E(\epsilon(\mu)\epsilon(\mu)^\top \mid X^*)$; 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 (μ,η) 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 $\mu_{n,globeff} \pm z_{\alpha/2} I^{-1/2}(\mu,\eta)/\sqrt{n}$ based on $\mu_{n,globeff}$ and the semiparametric variance bound $I^{-1}(\mu,\eta)$ will be considerably less than its nominal $(1-\alpha)$ level at laws (μ,η) at which $E(\epsilon(\mu)\epsilon(\mu)^\top \mid X^*)$ 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 $E(\epsilon(\mu)\epsilon(\mu)^\top \mid X^*)$ will be biased so that h_{opt} and its estimate will differ greatly but (ii) if a small bandwidth is used, the second-order $o_P(1/\sqrt{n})$ terms in the asymptotic linearity expansion $\mu_{n,globeff} - \mu = \frac{1}{n}\sum_{i=1}^n IC(Y_i) + o_P(1/\sqrt{n})$, where $IC(Y)$ denotes the influence curve of $\mu_{n,globeff}$, 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 curse of dimensionality appropriate (CODA) asymptotics that serves as a much better guide.

Under CODA asymptotics, an estimator $\mu_{n,globeff}$ of a one-dimensional parameter μ is defined to be globally CODA-efficient if $\mu_{n,globeff} \pm z_{\alpha/2} I^{-1/2}(\mu,\eta)/\sqrt{n}$ (or equivalently $\mu_{n,globeff} \pm z_{\alpha/2}\widetilde{I}^{-1/2}(\mu,\eta)/\sqrt{n}$, where $\widetilde{I}$ is a uniformly consistent estimator of I) is an asymptotic $(1-\alpha)$ confidence interval for μ uniformly over all laws (μ,η) allowed by the model. An estimator $\mu_{n,loceff}$ is locally CODA-efficient at a working submodel if (i) $\mu_{n,loceff} \pm z_{\alpha/2} I^{-1/2}(\mu,\eta)/\sqrt{n}$ is an asymptotic $(1-\alpha)$ confidence interval for μ uniformly over all laws (μ,η) in the submodel and (ii) $\mu_{n,loceff} \pm z_{\alpha/2}\sigma_n$ is an asymptotic $(1-\alpha)$ confidence interval for μ uniformly over all laws (μ,η), where σ_n is the nonparametric bootstrap estimator of the standard error of $\mu_{n,loceff}$ (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 (μ,η) in the working submodel. These definitions extend to a vector parameter μ by requiring that they hold for each one-dimensional linear combination μ of the components. Arguments similar to those in Robins and Ritov (1997) show that in the model with $E(\epsilon(\mu)\epsilon(\mu)^\top \mid X^*)$ unrestricted, except by continuity and a bound on its matrix norm, no globally efficient CODA estimators exist (owing to undercoverage under certain laws (μ,η) 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 $\mu_{n,loceff} \pm z_{\alpha/2}\sigma_n$ for μ 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 $\mu_{n,globeff}$ can be made globally CODA-efficient if we impose the additional assumption that $E(\epsilon(\mu)\epsilon(\mu)^\top \mid X^*)$ 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 $\mu_{n,globeff}$, there will exist laws (μ,η) satisfying the local smoothness assumption such that the coverage of $\mu_{n,globeff} \pm z_{\alpha/2} I^{-1/2}(\mu,\eta)/\sqrt{n}$ 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. □

1.2.3 Coarsening at random

The distribution of Y is indexed by the distribution F_X of the full data structure X and the conditional distribution $G(\cdot \mid X)$ 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 $G(\cdot \mid X)$ as the censoring or coarsening mechanism. If the censoring variable C is allowed to depend on unobserved components of X, then μ is typically not identifiable from the distribution of Y without additional strong untestable 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