Unbiased risk estimation via Gaussian data thinning with unknown variance

Xinyue Fu, Ronan Perry

July 27, 2026

1 Introduction

1.1 Risk minimization

A fundamental problem in statistics and machine learning is how to obtain valid inference on models selected using the data. One prominent example is the problem of estimating the risk of a model learned from the data. To formalize this idea, assume that we have n samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, each composed of a feature vector $X_i \in \mathbb{R}^p$ and outcome $Y_i \in \mathbb{R}$. Let $\mathcal{F}$ denote a set of possible predictors and for each predictor $f \in \mathcal{F}$ let $\ell_f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ be a loss function. The *risk* of a predictor $f \in \mathcal{F}$ is denoted by

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\ell_f(X_i, Y_i)\right].$$

When the samples are independent and identically distributed (i.i.d.), then $R_n(f) = \mathbb{E}\left[\ell_f(X_i, Y_i)\right]$. Since $R_n(f)$ is a population quantity, a common estimator is the empirical risk:

$$\hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell_f(X_i, Y_i).$$

A model is commonly learned so as to minimize the empirical risk. We call this model the empirical risk minimizer (ERM) solution:

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \hat{R}_n(f).$$

However, having learned $\hat{f}$ from the data, the empirical risk $\hat{R}_n(\hat{f})$ no longer becomes an unbiased estimator of the true risk $R_n(\hat{f})$ since $\hat{f}$ was selected precisely so as to do well on the samples used in the empirical risk calculation, so-called "double dipping".

1.2 Estimation after model selection via Gaussian data thinning

A popular approach for risk estimation is sample splitting: (i) a subset of the data are used to learn a model, and then (ii) the remaining samples are used to estimate the risk of this model. Because the sample splits are independent (assuming independent samples), the risk estimate is an unbiased estimator. Extensions such as cross-validation help to mitigate sensitivity due to the randomness of the splits. However, these estimators are for the risk of the model learned on a subset of the data, not the full model $\hat{f}$ learned on all of the data and which will be actually deployed.

Furthermore, sample splitting is not always applicable. In cases where the data are not independent, e.g., time-series, spatial, or graph-valued data, sample splitting is no longer meaningful. Similar issues arise when we care about the specific covariates X_i, e.g., in fixed-X regression. Lastly, when the sample size is low, especially relative to the dimensionality p of the covariates, sample splitting can lead to poor numerical behavior, e.g., in linear regression the matrix $X^\top X$ can become ill-conditioned.

An alternative to sample splitting is data thinning [Rasines and Young, 2023, Neufeld et al., 2024, Dharamshi et al., 2025, Leiner et al., 2025]. In the case of Gaussian data, data thinning works as follows. Suppose that $Y \sim \mathrm{N}(\mu, \sigma^2 I_n)$ where σ^2 is known. Then for any $\alpha > 0$ and $\omega \sim \mathrm{N}(0, \sigma^2 I_n)$,

$$Y^{\text{train}} := Y + \sqrt{\alpha}\omega \quad \text{and} \quad Y^{\text{test}} := Y - \omega/\sqrt{\alpha}$$

are independent random variables with mean μ and inflated variances. Why is this relevant? Y^{train} can be used to learn a model and Y^{test} can be used for independent risk estimation. The value of α trades off information, i.e., signal to noise ratio, just as in sample splitting you choose the splitting fraction. The caveat is that Y must be normally

distributed with a known variance.

To be more formal, we can use Y^{train} to learn the model

$$\hat{f}^{\text{train}} := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell_f \left(X_i, Y_i^{\text{train}} \right)$$

and estimate the risk using Y^{test}. Since Y^{train} is independent of Y^{test}, we will not incur bias due to reuse of the training data. However, this requires not only that the data be Gaussian, but also that the variance σ^2 is known.

2 Risk estimation with an unknown variance

A limitation of the standard data thinning framework discussed above is that it assumes a known variance and requires the data distribution and the constructed noise distribution to have identical variances. Specifically,

$$Y \sim \mathrm{N}\left(\mu, \sigma^2 I_n\right) \qquad \text{and} \qquad \omega \sim \mathrm{N}\left(0, \sigma_\omega^2 I_n\right),$$

where $\sigma^2 = \sigma_\omega^2$. Previous works [Tian, 2020, Oliveira et al., 2024, Liu et al., 2026] have used this strong assumption to (i) learn $\hat{f}^{\text{train}}$ using Y^{train} and (ii) estimate the prediction error with squared error loss:

$$R_n(\hat{f}^{\text{train}}) = \sum_{i=1}^{n} \mathbb{E}\left[\left(\tilde{Y}_i - \hat{f}^{\text{train}}(X_i) \right)^2 \right],$$

where the vector $\tilde{Y}$ is independent and identically distributed as Y. Liu et al. [2026] defines the risk estimator

$$\tilde{R}_n(\hat{f}^{\text{train}}) := \left\| Y^{\text{test}} - \hat{f}^{\text{train}}(X) \right\|_2^2 - \frac{1}{\alpha} \left\| \omega \right\|_2^2,$$

which is an unbiased estimator for $R_n(\hat{f}^{\text{train}})$. However, these works do not study the consequences when $\sigma_\omega^2 \neq \sigma^2$, e.g., when we use an estimator for σ^2.

The following result addresses the case of arbitrary σ_ω^2 in the case of linear models.

Lemma 1. *Suppose that there exists a positive semi-definite matrix H such that $\hat{f}(X) = HY$. Then*

$$R_n(\hat{f}) = \mathbb{E}\left[\tilde{R}_n(\hat{f}^{\text{train}})\right] - 2\left(\sigma_\omega^2 - \sigma^2\right)\text{tr}(H) - \alpha\sigma_\omega^2\,\text{tr}(H^2).$$

Lemma 1 implies that

$$\tilde{R}_n(\hat{f}^{\text{train}}) - 2\left(\sigma_\omega^2 - \sigma^2\right)\text{tr}(H) - \alpha\sigma_\omega^2\,\text{tr}(H^2)$$

is an unbiased estimator for the risk $R_n(\hat{f})$. The term $2\left(\sigma_\omega^2 - \sigma^2\right)\text{tr}(H)$ accounts for the bias due to dependence between Y^{train} and Y^{test}, while the term $\alpha\sigma_\omega^2\,\text{tr}(H^2)$ accounts for the bias due to learning $\hat{f}^{\text{train}}$ from noisy training data. We can see that when $\sigma_\omega^2 < \sigma^2$, $\tilde{R}_n(\hat{f}^{\text{train}})$ will underestimate the risk $R_n(\hat{f})$. This aligns with the result of Liu et al. [2026] that when $\sigma_\omega^2 = \sigma^2$ and $\alpha \to 0$, $\tilde{R}_n(\hat{f}^{\text{train}})$ approaches the desired estimand $R_n(\hat{f})$, recalling that $\hat{f}$ is learned using Y instead of Y^{train}.

3 Simulation studies

We conducted a simulation study to verify Lemma 1. Let $X \in \mathbb{R}^{n\times d}$ with $n = 100$, $d = 20$, and entries $X_{ij} \overset{iid}{\sim} \text{N}\left((0, \sigma^2)\right.$ where $\sigma^2 = 1$. We let $\beta^* \in \mathbb{R}^d$ with entries $\beta_i^* = 2^{-i+1}$ for $i \in [d]$, and sample n outcomes $Y = X\beta^* + \varepsilon$ where $\varepsilon_i \overset{iid}{\sim} \text{N}\left(0, 1\right)$. Let $\alpha = 1$. We sample $W \sim \text{N}\left(0, \sigma_\omega^2 I_n\right)$ for $\sigma_\omega^2 = \rho\sigma^2$ for $\rho \in \{0, 0.1, 0.5, 0.9, 1, 1.1, 2, 10\}$.

On the x-axis, we plot values of ρ, and on the y-axis, we plot the average of 1000 repetitions of (i) $R_n(\hat{f})$, (ii) the biased estimator $\tilde{R}_n(\hat{f}^{\text{train}})$, and (iii) our unbiased estimator (with standard errors).

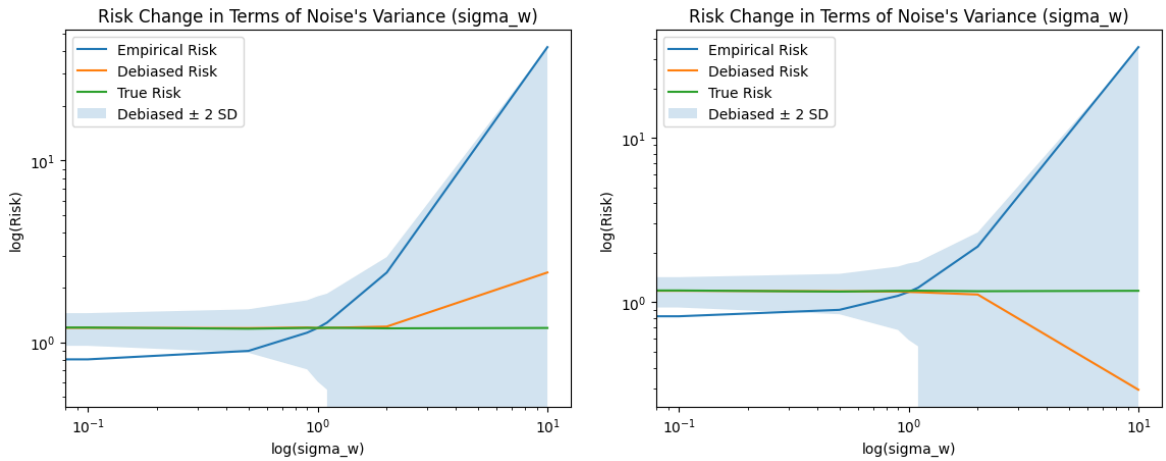

Figure 1: **Left:** OLS **Right:** Ridge regression. Comparison of empirical, debiased, and true risks under varying levels of thinning noise. The empirical risk exhibits increasing sensitivity to σ_w for both OLS and Ridge regression. While our unbiased estimator tracks the true risk reasonably well for smaller values of σ_w, its behavior diverges from the true risk as σ_w increases, particularly in the Ridge regression setting. The shaded region indicates ± 2 standard deviations of our estimator.

References

Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L. Gao, Daniela Witten, and Jacob Bien. Generalized Data Thinning Using Sufficient Statistics. *Journal of the American Statistical Association*, 120(549):511–523, January 2025. ISSN 0162-1459.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data Fission: Splitting a Single Data Point. *Journal of the American Statistical Association*, 120(549): 135–146, January 2025.

Sifan Liu, Snigdha Panigrahi, and Jake A Soloff. Cross-validation with antithetic gaussian randomization. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkag073, 2026.

Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data Thinning for Convolution-Closed Distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

Natalia L. Oliveira, Jing Lei, and Ryan J. Tibshirani. Unbiased risk estimation in the normal means problem via coupled bootstrap techniques. *Electronic Journal of Statistics*, 18(2):5405–5448, January 2024.

D García Rasines and G A Young. Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614, September 2023. ISSN 1464-3510.

Xiaoying Tian. Prediction error after model search. *The Annals of Statistics*, 48(2): 763–784, April 2020.

A Proof of Lemma 1

Proof. We will make use of the following fact.

Proposition 2. *If* $Z \sim \mathrm{N}(a, B)$*, then*

$$\begin{aligned}\mathbb{E}[\|Z\|_2^2] &= \mathrm{tr}(\mathrm{Var}[Z]) + a^\top a \\ &= \mathrm{tr}(B) + a^\top a\end{aligned}$$

According to Theorem 4.2 (Bias) from Liu et al. [2026]:

$$\begin{aligned}\mathbb{E}\left[\tilde{R}_n(\hat{f}^{\text{train}})\right] &= \mathbb{E}\left[\left\|Y^{\text{test}} - \hat{f}^{\text{train}}(X)\right\|_2^2 - \frac{1}{\alpha}\|\omega\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\left(Y - \frac{1}{\sqrt{\alpha}}\omega\right) - \hat{f}^{\text{train}}(X)\right\|_2^2 - \frac{1}{\alpha}\|\omega\|_2^2\right] \\ &= \mathbb{E}\left[\left\|Y - \frac{1}{\sqrt{\alpha}}\omega\right\|_2^2 - 2\left(Y - \frac{1}{\sqrt{\alpha}}\omega\right)^\top \hat{f}^{\text{train}}(X)\right. \\ &\qquad \left. + \left\|\hat{f}^{\text{train}}(X)\right\|_2^2 - \frac{1}{\alpha}\|\omega\|_2^2\right] \\ &= \mathbb{E}\left[\left\|Y - \frac{1}{\sqrt{\alpha}}\omega\right\|_2^2\right] - 2\,\mathbb{E}\left[\left(Y - \frac{1}{\sqrt{\alpha}}\omega\right)^\top HY^{\text{train}}\right] \\ &\quad + \mathbb{E}\left[\left\|HY^{\text{train}}\right\|_2^2\right] - \frac{1}{\alpha}\mathbb{E}\left[\|\omega\|_2^2\right]\end{aligned}$$

Term 1: Addressing the first term, since $Y \sim \mathrm{N}(\mu, \sigma^2 I_n)$ and $\mathbb{E}[\omega] = 0$:,

$$\begin{aligned}\mathbb{E}\left[\left\|Y - \frac{1}{\sqrt{\alpha}}\omega\right\|_2^2\right] &= \mathbb{E}\left[\left(Y - \frac{1}{\sqrt{\alpha}}\omega\right)^\top \left(Y - \frac{1}{\sqrt{\alpha}}\omega\right)\right] \\ &= \mathbb{E}\left[Y^\top Y - 2\frac{1}{\sqrt{\alpha}}Y^\top\omega + \left(\frac{1}{\sqrt{\alpha}}\omega\right)^2\right] \\ &= \mathbb{E}\left[\|Y\|_2^2\right] - 2\frac{1}{\sqrt{\alpha}}\mathbb{E}[Y^\top]\mathbb{E}[\omega] + \mathbb{E}\left[\left(\frac{1}{\sqrt{\alpha}}\right)^2\|\omega\|_2^2\right] \\ &= \mathbb{E}\left[\|Y\|_2^2\right] + \frac{1}{\alpha}\mathbb{E}\left[\|\omega\|_2^2\right] \\ &= n\sigma^2 + \|\mu\|_2^2 + \frac{1}{\alpha}\mathbb{E}\left[\|\omega\|_2^2\right].\end{aligned}$$

Term 2: Since $H \in \mathbb{R}^{n\times n}$ is positive semi-definite, by Proposition 2 $\mathbb{E}[\omega^\top H\omega] = \sigma_\omega^2\,\mathrm{tr}(H)$ and $\mathbb{E}[Y^\top HY] = \sigma^2\,\mathrm{tr}(H) + \mu^\top H\mu$. Thus,

$$
\begin{aligned}
\mathbb{E}\left[\left(Y-\frac{1}{\sqrt{\alpha}}\omega\right)^{\top}\hat{f}^{\text{train}}(X)\right] &= \mathbb{E}\left[\left(Y^{\top}-\frac{1}{\sqrt{\alpha}}\omega^{\top}\right)\hat{f}^{\text{train}}(X)\right] \\
&= \mathbb{E}\left[Y^{\top}\hat{f}^{\text{train}}(X)-\frac{1}{\sqrt{\alpha}}\omega^{\top}\hat{f}^{\text{train}}(X)\right] \\
&= \mathbb{E}\left[Y^{\top}H\left(Y+\sqrt{\alpha}\omega\right)\right]-\frac{1}{\sqrt{\alpha}}\mathbb{E}\left[\omega^{\top}H\left(Y+\sqrt{\alpha}\omega\right)\right] \\
&= \mathbb{E}\left[Y^{\top}HY+\sqrt{\alpha}\,Y^{\top}H\omega\right]-\frac{1}{\sqrt{\alpha}}\mathbb{E}\left[\omega^{\top}HY+\sqrt{\alpha}\,\omega^{\top}H\omega\right] \\
&= \mathbb{E}\left[Y^{\top}HY\right]+\sqrt{\alpha}\,\mathbb{E}\left[Y^{\top}H\omega\right] \\
&\quad -\frac{1}{\sqrt{\alpha}}\mathbb{E}\left[\omega^{\top}HY\right]-\mathbb{E}\left[\omega^{\top}H\omega\right] \\
&= \mathbb{E}\left[Y^{\top}HY\right]+\sqrt{\alpha}\,\mathbb{E}[Y^{\top}]H\mathbb{E}[\omega] \\
&\quad -\frac{1}{\sqrt{\alpha}}\mathbb{E}[\omega^{\top}]H\mathbb{E}[Y]-\mathbb{E}\left[\omega^{\top}H\omega\right] \\
&= \mathbb{E}\left[Y^{\top}HY\right]-\mathbb{E}\left[\omega^{\top}H\omega\right] \\
&= \sigma^2\operatorname{tr}(H)+\mu^{\top}H\mu-\sigma_{\omega}^2\operatorname{tr}(H) \\
&= (\sigma^2-\sigma_{\omega}^2)\operatorname{tr}(H)+\mu^{\top}H\mu
\end{aligned}
$$

Term 3: Lastly, addressing the third term,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{f}^{\text{train}}(X)\right\|_2^2\right] &= \mathbb{E}\left[\left\|H\left(Y+\sqrt{\alpha}\omega\right)\right\|_2^2\right] \\
&= \operatorname{tr}\left(\operatorname{Var}\left[H\left(Y+\sqrt{\alpha}\omega\right)\right]\right)+(H\mu)^{\top}(H\mu) \\
&= \operatorname{tr}\left((\sigma^2+\alpha\sigma_{\omega}^2)H^2\right)+\mu^{\top}H^2\mu \\
&= (\sigma^2+\alpha\sigma_{\omega}^2)\operatorname{tr}(H^2)+\mu^{\top}H^2\mu.
\end{aligned}
$$

Combining: Combining all terms, we find that

$$
\mathbb{E}\left[\tilde{R}_n(\hat{f}^{\text{train}})\right] = (\sigma^2+\alpha\sigma_{\omega}^2)\operatorname{tr}(H^2)+\mu^{\top}H^2\mu-2(\sigma^2-\sigma_{\omega}^2)\operatorname{tr}(H)-2\mu^{\top}H\mu+n\sigma^2+\|\mu\|_2^2
$$

We now expand the true risk $R_n(\hat{f})$:

$$\begin{aligned}
R_n(\hat{f}) &= \mathbb{E}\left[\left\|\hat{f}(X) - \widetilde{Y}\right\|_2^2\right] \\
&= \mathbb{E}\left[\|\hat{f}(X)\|_2^2 - 2\hat{f}(X)^\top \widetilde{Y} + \|\widetilde{Y}\|_2^2\right] \\
&= \mathbb{E}\left[\|\hat{f}(X)\|_2^2\right] - 2\,\mathbb{E}[\hat{f}(X)]^\top \mathbb{E}[\widetilde{Y}] + \mathbb{E}\left[\|\widetilde{Y}\|_2^2\right] \\
&= \operatorname{tr}(\sigma^2 H^2) + (H\mu)^\top (H\mu) - 2(H\mu)^\top \mu + \left[\operatorname{tr}(\sigma^2 I_n) + \mu^\top \mu\right] \\
&= \sigma^2 \operatorname{tr}(H^2) + \mu^\top H^2 \mu - 2\mu^\top H\mu + n\sigma^2 + \|\mu\|_2^2.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\tilde{R}_n(\hat{f}^{\text{train}})\right] = R_n(\hat{f}) + \underbrace{2(\sigma_\omega^2 - \sigma^2)\operatorname{tr}(H)}_{\text{Bias due to incorrect variance}} + \underbrace{\alpha\sigma_\omega^2 \operatorname{tr}(H^2)}_{\text{Bias due to noisy training data}}.$$

□