J. Sophia Wang and P. M. Aronow. Supporting Information for Listwise Deletion in High Dimensions.

Supporting Information A. Numerical Illustration of Theoretical Findings.

Here we demonstrate the finite-n implications of our findings. In the following figures, we plot figures for different values of an upper bound on the conditional probability of each data entry being observed q_*, the number of observations (rows) n, the number of variables (columns) k and a sharp lower bound on p_{all} denoted $\underline{p_{all}} = (1 - q_*^{f(n)})^n$. $\underline{p_{all}}$ is the lowest possible ("best-case") probability that listwise deletion removes all data.

Figure 3 demonstrates how listwise deletion asymptotically removes all data. Figure 3 plots the sharp lower bound $\underline{p_{all}}$ for the probability that listwise deletion removes all data against the number of variables k in three different settings for $n = 100, 1000, 10000$ in each subfigure. By Lemma 3, we can compute $\underline{p_{all}} = (1 - q_*^k)^n$.

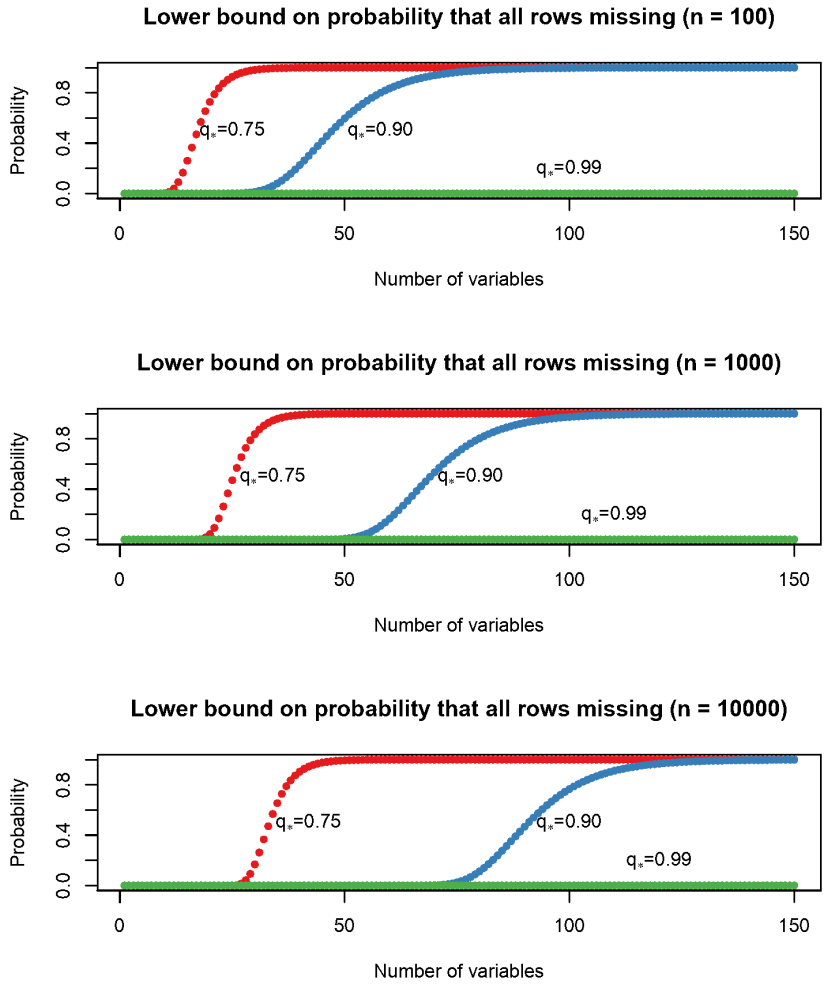


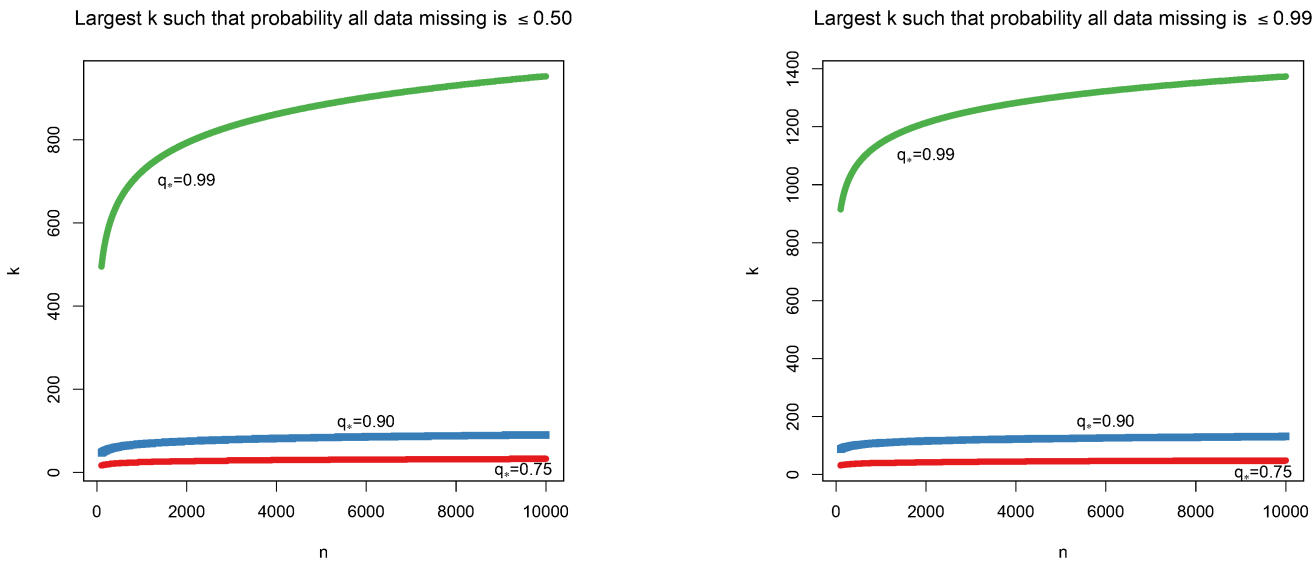
Figure 3. Lower bound of probability that all rows missing ($\underline{p_{all}}$) plotted against values of n,k and q_*.

In each subfigure, we simultaneously consider three different values for the upper bound of the conditional

probability q_* defined in Assumption 2: the red curves represent $\underline{p_{all}}$ calculated with $q_* = 0.75$, the blue curves represent $\underline{p_{all}}$ calculated with $q_* = 0.90$, and the green curves represent $\underline{p_{all}}$ calculated with $q_* = 0.99$. We see that the rate of $\underline{p_{all}}$ converges to 1 as k gets large, which is faster when q gets smaller. However, the rate of convergence heavily depends on q_*. When $q_* = 0.75$ (i.e., each variable has at least a 25% chance of idiosyncratic missingness), the lower bound $\underline{p_{all}}$ is extremely close to 1 even when $n = 10,000$. However, when the upper bound q_* is as large as 0.99, $\underline{p_{all}}$ is essentially zero when $n = 100$ and $k = 150$, reflecting the fact that the probability of idiosyncratic missingness is essential in determining the properties of listwise deletion.

Figure 4 illustrates, for a given n, q_*, how large k can be while still ensuring that $p_{all} \leq 0.5, 0.99$. We compute this using the result from Lemma 3, $\underline{p_{all}} = (1 - q_*^k)^n$. Since $(1 - q_*^k)^n$ is strictly increasing in k, solving for equality $p_{all} = (1 - q_*^k)^n$ we will get the smallest possible k for each $\underline{p_{all}}$ that $k = \left\lfloor \frac{\log(1-\underline{p_{all}}^{1/n})}{\log(q_*)} \right\rfloor$. We present two subfigures: with $\underline{p_{all}} = 0.5$ for the first subfigure and $\underline{p_{all}} = 0.99$ for the second subfigure, and we plot the k against the n for three different upper bounds $q_* = 0.75, 0.90, 0.99$.

With more missingness, $q_* = 0.75, 0.90$, even relatively small k can yield missingness of $\underline{p_{all}}$. For example, even with $n = 10,000$ and $q_* = 0.75$ we need only $k = 33$ to have a 50% probability that all rows will be missing. However, when missingness is very low, k needs to be very large to cause all data to be missing. For example, with $n = 10,000$ and $q_* = 0.99$, we need $k = 952$ to have a 50% probability that all rows will be missing.



(a) $p_{all} \leq 0.5$ **(b)** $p_{all} \leq 0.99$

Figure 4. Largest k such that $p_{all} \leq 0.5, 0.99$ plotted against n and q_*.

Our final numerical illustration considers an upper bound on the expected proportion of observations that are missing, $1 - q_*^k$, which does not depend on n. Figure 5 plots the expected proportion of data missing versus the number of variables k. We see the same qualititative relationship as before — as the number of

variables increases, we have a very quick decline in the proportion of usable data. In comparison to Figure 3, the expected proportion of data missing tends faster to 1 for each q_* considered as k gets large, as it is equivalent to the special case for $\underline{p_{all}}$ when $n = 1$.

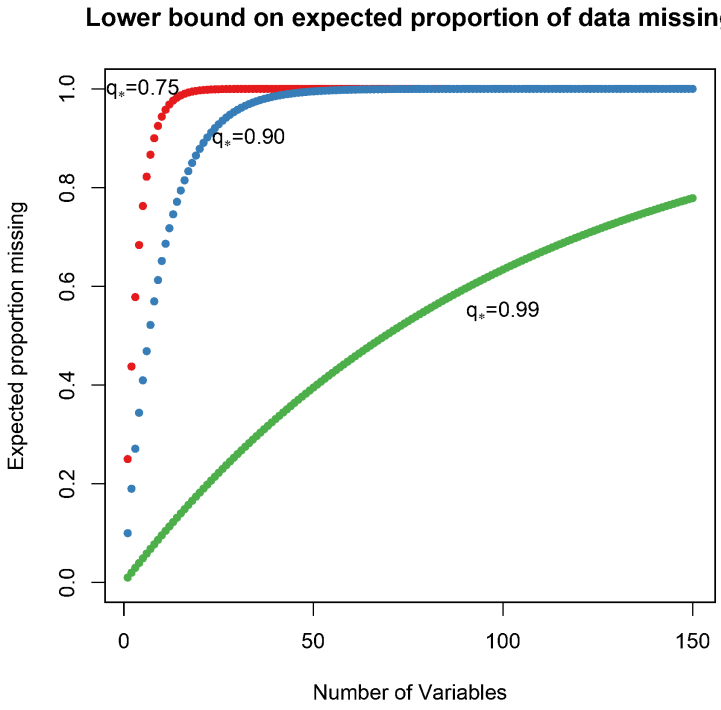


Figure 5. Lower bound of expected proportion of all rows missing ($\underline{p_{all}}$) plotted against values of n,k and q_*.

Supporting Information B. Asymptotics in the Number of Groups of Variables.

In this section, we provide a formal exposition of how our results can generalize to the case where we have idiosyncratic missingness with respect to groups of variables, rather than each specific variable. The language is largely duplicative of the language in Section 2 of the main text; however it makes explicit the direct manner in which the result can generalize.

Let n be the number of observations. Let g be the number of variable groups in the dataset, within which all observations share an identical missingness pattern. We let M_{ij} be a random indicator variable for whether or not the jth group in the ith row is missing. We use one indicator for each group due to the shared missingness. Similar to the previous setting, we use $\mathbf{M}_{ij}$ to represent the random vector collecting the missingness indicators up to variable j, $(M_{i1}, M_{i2}, \ldots, M_{ij})$. Let k be the number of variables in the dataset. Note that, by construction, we know that $k \geq g$ since groups contain at least one variable.

We will restate Assumption 1 and 2 in the group settings in Assumption 6 and 7 such that there is mutual independence of missingness across rows as well as the conditional probability that an observation is missing being bounded away from zero.

Assumption 6. *All rows of the data* $\Big((M_{11}, ..., M_{1g}), ..., (M_{n1}, ..., M_{ng})\Big)$ *are mutually independent.*

Assumption 7. *There exists a* $q_* \in [0, 1)$ *such that for all* i*,*

- $\Pr(M_{i1} = 0) \leq q_*$
- $\Pr(M_{ij} = 0 | \mathbf{M}_{i(j-1)} = \mathbf{0}) \leq q_*$*, for all* $j \in \{2, \ldots, g\}$ *such that* $\Pr(\mathbf{M}_{i(j-1)} = \mathbf{0}) > 0$

Then we can obtain a group version of Lemma 3 (Lemma 8) following similar steps.

Lemma 8. *Under Assumptions 6 and 7, the probability that listwise deletion removes all rows is* $p_{all} \geq (1 - q_*^g)^n$.

Proof of Lemma 8: Similar to the proof of Lemma 3, we will still consider two cases for q_*. Suppose $q_* = 0$. Since this entails that all groups of variables are completely missing, $p_{all} = 1 = (1 - q_*^g)^n$. For the second case suppose $q_* \in (0, 1)$. By the group independence assumption, $p_{all} = \Pi_{i=1}^n (1 - \Pr(\mathbf{M}_{ig} = \mathbf{0}))$. Denote $q_{ij} = \Pr(M_{ij} = 0 | \mathbf{M}_{i(j-1)} = \mathbf{0})$ if $j > 1$, else $q_{ij} = \Pr(M_{ij} = 0)$. By Assumption 7, $q_{ij} \leq q_*$, for all $j \in \{1, 2, \ldots, g\}$. By the chain rule of conditional probability, $\Pr(\mathbf{M}_{ig} = \mathbf{0}) = q_{i1} q_{i2} \cdots q_{ig}$. This means that the probability of a single observation containing at least one missing entry is $(1 - q_{i1} q_{i2} \cdots q_{ig})$. Since $q_* \geq q_{ij}$ for all $j \in \{1, 2, \ldots, g\}$, $q_*^g \geq q_{i1} q_{i2} \cdots q_{ig}$. Thus $(1 - q_*^g) \leq (1 - q_{i1} q_{i2} \cdots q_{ig})$. Thus $(1 - q_*^g)^n$ is a lower bound for the probability of all n observations each containing at least one missing entry. $\square$

Similarly, we will embed the problem into a sequence $g_n = l(n)$, where l has range over the natural numbers, and allow $M_{ij,n}$ and $q_{ij,n}$ to vary at each n. We omit the n notation for simplicity. Hence we have the third assumption in the group setting that g grows superlogarithmically in n. We discussed the interpretation of this assumption in the variable setting extensively in the Theory section, so here we will only present the assumption and the group version of Proposition 5 in Proposition 10, as well as a proof for Proposition 10.

Assumption 9. *The number of groups of covariates grows superlogarithmically in* n*, so that* $\lim_{n \to \infty} \frac{l(n)}{log(n)} = \infty$.

Proposition 10. *Under Assumptions 6, 7 and 9,* $\lim_{n \to \infty} p_{all} = 1$.

Proof of Proposition 10: First we will show that $\lim_{n \to \infty} n q_*^{l(n)} = 0$ (in asymptotic shorthand notation, $q_*^{l(n)} = o(n^{-1})$). Note that

$$\lim_{n \to \infty} n q_*^{l(n)} = \lim_{n \to \infty} e^{\log n q_*^{l(n)}} = \lim_{n \to \infty} e^{\log n + l(n) \log q_*}.$$

Since $q_* \in (0, 1)$, $\log q_* < 0$. Since $l(n) = \omega(\log n)$, the sequence $\log n + l(n) \log q_*$ diverges to negative infinity, and so

$$\lim_{n \to \infty} e^{\log n + l(n) \log q_*} = 0 = \lim_{n \to \infty} n q_*^{l(n)}.$$

Since $q_* \in (0, 1)$ and $k = l(n) \geq_* 1$, $-q_*^{l(n)} > -1$ and $1 - q_*^{l(n)} \leq 1$. By Bernoulli's Inequality, since $n \in \mathbb{N}, (1 - q_*^{l(n)})^n \geq_* 1 + n(-q_*^{l(n)}) = 1 - n q_*^{l(n)}$. Thus $1 - n q_*^{l(n)} \leq (1 - q_*^{l(n)})^n \leq 1$ in the common

domain $n \in \mathbb{N}$. Since $\lim_{n\to\infty} 1 = 1$ and $\lim_{n\to\infty} 1 - nq_*^{l(n)} = 1 - \lim_{n\to\infty} nq_*^{l(n)} = 1$, by the Squeeze Theorem,

$$\lim_{n\to\infty} (1 - q_*^{l(n)})^n = 1.$$

Then, since $\forall n, (1 - q_*^{l(n)})^n \leq p_{all} \leq 1$, we have $\lim_{n\to\infty} p_{all} = 1$, again by the Squeeze Theorem. $\square$