BALANCED SUBSAMPLING FOR BIG DATA WITH CATEGORICAL PREDICTORS

Lin Wang

Purdue University

Supplementary Material

S1 Notations and Orthonormal Contrasts

Recall that $\mathcal{X}$ denotes the set of all possible level combinations of predictors, that is, $\mathcal{X} = \{x = (x_1, \ldots, x_p) : x_j = 1, \ldots, q_j, j = 1, \ldots, p\}$. Let $\mathcal{N} = \#\mathcal{X} = \prod_{j=1}^{p} q_j$. Let $\mathcal{Z}$ be the matrix of dummy variables for $\mathcal{X}$ and C be the coded matrix for $\mathcal{X}$ via orthonormal contrasts (Chen and Tang, 2022; Wang and Xu, 2022) with $C^T C = \mathcal{N} I$, where I is a conformable identity matrix. Then there exists a transformation matrix P such that $\mathcal{Z} = CP$. Because both $\mathcal{Z}$ and C have full column ranks, so P is nonsingular. Clearly, rows of Z_s come from rows of $\mathcal{Z}$, so $Z_s = C_s P$ and $M_s = P^T C_s^T C_s P$, where rows of C_s are from the corresponding rows of C.

Common orthonormal contrasts include (normalized) Helmert contrasts

(Chambers and Hastie, 2017), orthogonal polynomial contrasts (Wang and Xu, 2022), and complex contrasts Xu and Wu (2001). For example, the Helmert contrasts are used to contrast the second level with the first, the third with the average of the first two, and so on. For the jth predictor with q_j levels, the lth contrast ($l = 1, \ldots, q_j - 1$) at level u is

$$c_{jl}(u) = \sqrt{\frac{q_j}{l + l^2}} \cdot a_{jl}(u), \text{ where } a_{jl}(u) = \begin{cases} -1, & \text{for } u < l \\ l, & \text{for } u = l \\ 0, & \text{otherwise} \end{cases} \tag{S1.1}$$

Below is an example of the $a_{jl}(u)$ for a predictor with $q_j = 5$ levels, where the lth column lists the lth contrast at $u = 1, \ldots, 5$ levels:

$$\begin{pmatrix} -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$$

S2 Proof of Theorem 1

Since $M_s = P^T C_s^T C_s P$, then

$$\lambda_{\min}(M_s) = \lambda_{\min}(P^T C_s^T C_s P) \geq \lambda_{\min}(P^T P)\lambda_{\min}(C_s^T C_s) = \nu \lambda_{\min}(C_s^T C_s), \tag{S2.1}$$

where $\nu = \lambda_{\min}(P^TP) > 0$. Let $\tilde{C} = I - C_s^TC_s/n = C^TC/\mathcal{N} - C_s^TC_s/n$. Because each level u for the jth predictor appears $\mathcal{N}/q_j$ times in $\mathcal{X}$ and $n_j(u)$ times in X_s, then

$$\tilde{C} = \begin{pmatrix} 0 & C_1^T & C_2^T & \cdots & C_p^T \\ C_1 & C_{11} & C_{12}^T & \cdots & C_{1p}^T \\ C_2 & C_{12} & C_{22} & \cdots & C_{2p}^T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_p & C_{1p} & C_{2p} & \cdots & C_{pp} \end{pmatrix}$$

where C_j is a (q_j-1)-vector with the lth entry $\sum_{u=1}^{q_j}[q_j^{-1} - n_j(u)/n]c_{jl}(u) = \sum_{u=1}^{q_j}[q_j^{-1} - n_j(u)/n]c_{jl}(u)c_{j0}(u)$ with $c_{j0}(u) = 1$, C_{jj} is a $(q_j-1)\times(q_j-1)$ matrix with the (l,m)th entry $\sum_{u=1}^{q_j}[q_j^{-1} - n_j(u)/n]c_{jl}(u)c_{jm}(u)$, C_{jk} is a $(q_j-1)\times(q_k-1)$ matrix with the (l,m)th entry $\sum_{u=1}^{q_j}\sum_{v=1}^{q_k}[(q_jq_k)^{-1} - n_{jk}(u,v)/n]c_{jl}(u)c_{km}(v)$, and $c_{jl}(u)$ is the coded value for the lth contrast of the jth predictor at level u. For example, for the Helmert contrast, $c_{jl}(u)$ is given in (S1.1). We have

$$\lambda_{\max}(\tilde{C}) \le \|\tilde{C}\|_F = \sqrt{f_1 + f_2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm and

$$\begin{aligned} f_1 &= \sum_{j=1}^{p}\sum_{l=0}^{q_j-1}\sum_{m=0}^{q_j-1}\sum_{u=1}^{q_j}\left[\frac{1}{q_j} - \frac{n_j(u)}{n}\right]^2 c_{jl}^2(u)c_{jm}^2(u) \\ f_2 &= \sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p}\sum_{l=1}^{q_j-1}\sum_{m=1}^{q_k-1}\sum_{u=1}^{q_j}\sum_{v=1}^{q_k}\left[\frac{1}{q_jq_k} - \frac{n_{jk}(u,v)}{n}\right]^2 c_{jl}^2(u)c_{km}^2(v). \end{aligned}$$

Because $\sum_{l=0}^{q_j-1}\sum_{m=0}^{q_j-1} c_{jl}^2(u)c_{jm}^2(u) \le q_j^2$ and $\sum_{l=1}^{q_j-1}\sum_{m=1}^{q_k-1} c_{jl}^2(u)c_{km}^2(v) \le q_jq_k$ for any j,k and u,v, then

$$\begin{aligned}\lambda_{\max}(\tilde{C}) &\le \sqrt{\sum_{j=1}^{p}\sum_{u=1}^{q_j} q_j^2\left[\frac{1}{q_j}-\frac{n_j(u)}{n}\right]^2+\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p}\sum_{u=1}^{q_j}\sum_{v=1}^{q_k} q_jq_k\left[\frac{1}{q_jq_k}-\frac{n_{jk}(u,v)}{n}\right]^2} \\ &= f(X_s).\end{aligned}$$

Since $C_s^TC_s = n(I-\tilde{C})$, $\lambda_{\min}(C_s^TC_s) = n(1-\lambda_{\max}(\tilde{C})) \ge n(1-f(X_s))$. Then by (S2.1), $\lambda_{\min}(M_s) \ge n\nu(1-f(X_s))$.

S3 Proof of Theorem 2

Since $M_s = P^TC_s^TC_sP$, where P is a $Q\times Q$ transformation matrix with $Q = 1+\sum_{j=1}^{p}(q_j-1)$, then

$$\det(M_s) = \det(P^TC_s^TC_sP) = \det(P)^2\det(C_s^TC_s).$$

Because P is independent from the selection of Z_s, we only need to consider $\det(C_s^T C_s)$:

$$\begin{aligned}
\det(C_s^T C_s) &= \prod_{j=1}^{Q} \lambda_j(C_s^T C_s) \\
&\leq \left\{ \frac{\sum_{j=1}^{Q} \lambda_j(C_s^T C_s)}{Q} \right\}^Q \\
&= \left\{ \frac{\mathrm{tr}(C_s^T C_s)}{Q} \right\}^Q \qquad \text{(S3.1)} \\
&= \left\{ \frac{nQ}{Q} \right\}^Q \qquad \text{(S3.2)} \\
&= n^Q,
\end{aligned}$$

where $\lambda_j(C_s^T C_s)$'s for $j = 0, 1, \ldots, Q$ are eigenvalues of $C_s^T C_s$, and $\mathrm{tr}(C_s^T C_s) = \mathrm{tr}(C_s C_s^T) = nQ$ because rows of C_s are orthonormal. If $f(X_s) = 0$, X_s forms an orthogonal array and $C_s^T C_s = nI$, then $\det(C_s^T C_s) = n^Q$. This completes the proof.

S4 Proof of Theorem 3

We have

$$E[(Y - z^T\hat{\beta}_s)^2 | X_s] = E[(Y - z^T\beta)^2] + E[(z^T\beta - z^T\hat{\beta}_s)^2 | X_s] = \sigma^2(1 + z^T M_s^{-1} z)$$

and

$$\begin{aligned}\sum_{x\in\mathcal{X}} z^T M_s^{-1} z &= \mathrm{tr}(\mathcal{Z} M_s^{-1}\mathcal{Z}^T) = \mathrm{tr}\{\mathcal{Z}^T\mathcal{Z} M_s^{-1}\} = \mathrm{tr}\{P^T C^T C P (P^T C_s^T C_s P)^{-1}\} \\ &= \mathrm{tr}\{C^T C (C_s^T C_s)^{-1}\} = \mathcal{N}\mathrm{tr}\{(C_s^T C_s)^{-1}\} \geq \mathcal{N}Q^2/\mathrm{tr}(C_s^T C_s) = \mathcal{N}Q/n,\end{aligned}$$

where the last equation holds because $\mathrm{tr}(C_s^T C_s) = nQ$ following (S3.1) and (S3.2). Therefore, $\max_{x\in\mathcal{X}} z^T M_s^{-1} z \geq Q/n$ and $\max_{x\in\mathcal{X}} E[(Y - z^T\hat{\beta}_s)^2|X_s] \geq \sigma^2(1+Q/n)$. On the other hand, when $f(X_s) = 0$, X_s is balanced, and then for any z, $z^T M_s^{-1} z = z^T P^T P z/n = \|Pz\|_2^2/n$. Note that Pz is a row vector of C. The sum of squares of the ith row of C is $(1 + \sum_{j=1}^{p}\sum_{l=1}^{q_j-1} c_{i,jl}^2)/n = (1 + \sum_{j=1}^{p}(q_j - 1))/n = Q/n$. Therefore, $z^T M_s^{-1} z = Q/n$ and $E[(Y - z^T\hat{\beta}_s)^2|X_s] = \sigma^2(1 + Q/n)$. This completes the proof.

S5 Proof of Theorem 4

For a subsample $X_s = (x_{ij}^*)$, it can be verified that $\sum_{u=1}^{q_j} n_j(u) = n$, $\sum_{u=1}^{q_j}\sum_{v=1}^{q_k} n_{jk}(u,v) = n$, and $\sum_{i=1}^{n}\sum_{l=1}^{n}\delta_1(x_{ij}^*, x_{lj}^*)\delta_1(x_{ik}^*, x_{lk}^*) = \sum_{u=1}^{q_j}\sum_{v=1}^{q_k} n_{jk}(u,v)^2$

for any $j, k = 1, \ldots, p$. Then

$$
\begin{aligned}
& f^2(X_s) \\
= & \sum_{j=1}^{p}\sum_{u=1}^{q_j}\left(1-\frac{2q_j n_j(u)}{n}+\frac{q_j^2 n_j(u)^2}{n^2}\right)+\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p}\sum_{u=1}^{q_j}\sum_{v=1}^{q_k}\left(\frac{1}{q_j q_k}-\frac{2n_{jk}(u,v)}{n}+\frac{q_j q_k n_{jk}(u,v)^2}{n^2}\right) \\
= & -\sum_{j=1}^{p} q_j + n^{-2}\sum_{j=1}^{p} q_j^2\left[\sum_{u=1}^{q_j} n_j(u)^2\right] - p(p-1) + n^{-2}\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} q_j q_k\left[\sum_{u=1}^{q_j}\sum_{v=1}^{q_k} n_{jk}(u,v)^2\right] \\
= & n^{-2}\sum_{j=1}^{p}\sum_{k=1}^{p} q_j q_k\left[\sum_{u=1}^{q_j}\sum_{v=1}^{q_k} n_{jk}(u,v)^2\right] - \sum_{j=1}^{p} q_j - p(p-1) \\
= & n^{-2}\sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p} q_j q_k \delta_1(x_{ij}^*, x_{lj}^*)\delta_1(x_{ik}^*, x_{lk}^*) - \sum_{j=1}^{p} q_j - p(p-1) \\
= & n^{-2}\sum_{i=1}^{n}\sum_{l=1}^{n}[\delta(x_i^*, x_l^*)]^2 - \sum_{j=1}^{p} q_j - p(p-1) \\
= & 2n^{-2}\sum_{1\leq i<l\leq n}[\delta(x_i^*, x_l^*)]^2 + C.
\end{aligned}
$$

S6 Computing complexity and time

The computational complexity of Algorithm 1 is $O(Npn)$. The computational complexity of IBOSS is $O(N(\sum_{j=1}^{p} q_j)) = O(Np\bar{q})$, where $\bar{q}$ represents the average number of levels of predictors. For LEV, we use a fast Singular Value Decomposition method implemented in the R package "corpcor" to accelerate LEV, so the complexity is also $O(N(\sum_{j=1}^{p} q_j)) = O(Np\bar{q})$.

Tables S6.1 and S6.2 show the computing time for the subsampling process in the setting of the simulation studies. All computations are carried

out on a laptop running Windows 11 Pro with an Intel Core i7-12700H processor and 32GB memory. When $n = 500$, the running time of balanced subsampling is comparable to that of fast LEV, with both being relatively faster than IBOSS. When $n = 2000$, the running time of balanced subsampling increases to four times that of $n = 500$, making its running time around four times that of fast LEV and 2.5 times that of IBOSS.

Given the constraints of limited resources for labeling data (observing the response), the subsample size n is typically not large. In such scenarios, balanced subsampling exhibits comparable computational time to other subsampling methods, making it viable for handling large datasets.

Table S6.3 presents the computing times for the real data application. With the real data featuring more levels and a larger $\bar{q}$, the advantage of balanced subsampling in terms of computing time is evident.

However, the primary objective of this paper is to identify an optimal subsampling approach for measurement-constrained regression, with the goal of achieving superior estimation and predictive performance. While computational efficiency is undoubtedly important, it takes a secondary role to the primary concern of achieving optimal performance.

Table S6.1: Running time (in seconds) of subsampling methods when $p = 20$, $q = 2, \ldots, (p+1)$, and the subsample size is $n = 500$.

N	UNI	IBOSS	LEV	Balanced
10^4	0	1.63	1.09	1.19
10^5	0	20.51	11.95	13.88
10^6	0	207.63	124.20	131.40

Table S6.2: Running time (in seconds) of subsampling methods when $p = 20$, $q = 2, \ldots, (p+1)$, and the subsample size is $n = 2000$.

N	UNI	IBOSS	LEV	Balanced
10^4	4×10^{-4}	1.56	1.05	4.21
10^5	8×10^{-4}	20.96	12.68	56.31
10^6	0.002	228.88	131.90	563.68

Table S6.3: Running time (in seconds) of the subsampling process for the real data.

n	UNI	IBOSS	LEV	Balanced
500	4×10^{-4}	26.39	26.34	4.24
2000	0.0012	40.53	31.41	21.34

References

Chambers, J. M. and T. J. Hastie (2017). Statistical models. In *Statistical models in S*, pp. 13–44. Routledge.

Chen, G. and B. Tang (2022). A study of orthogonal array-based designs under a broad class of space-filling criteria. *The Annals of Statistics 50*(5), 2925–2949.

Wang, L. and H. Xu (2022). A class of multilevel nonregular designs for studying quantitative factors. *Statistica Sinica 32*(2), 825–845.

Xu, H. and C. F. J. Wu (2001). Generalized minimum aberration for asymmetrical fractional factorial designs. *Annals of Statistics 29*(4), 1066–1077.