It's All In The (Exponential) Family: An Equivalence Between Maximum Likelihood Estimation and Control Variates For Sketching Algorithms

Keegan Kang
Department of Mathematics and Statistics, Bucknell University, USA

Kerong Wang
Department of Mathematics and Statistics, Bucknell University, USA

Ding Zhang
School of Data Science, University of Virginia, USA

Rameshwar Pratap
Department of CSE, IIT Hyderabad, India

Bhisham Dev Verma
Department of Computer Science, Wake Forest University, USA

Benedict H. W. Wong
Institute for Infocomm Research, (I^2R), A*STAR, Singapore

Abstract

Maximum likelihood estimators (MLE) and control variate estimators (CVE) have been used in conjunction with known information across sketching algorithms and applications in machine learning. We prove that under certain conditions in an exponential family, an optimal CVE will achieve the same asymptotic variance as the MLE, giving a fixed point algorithm for the MLE. Experiments show the fixed point algorithm is faster and numerically stable compared to other root finding algorithms for the MLE for the bivariate Normal distribution, and we expect this to hold across distributions satisfying these conditions. We show how this algorithm leads to reproducibility for algorithms using MLE / CVE, and demonstrate how the algorithm leads to finding the MLE when the CV weights are known.

1 INTRODUCTION

Many applications of estimation via sketching algorithms use information about known parameters, such as marginal information, in conjunction with maximum likelihood estimation (Shao, 2003) or control variates (Lavenberg and Welch, 1981). Examples range from estimating inner products (Li et al., 2006) or angles (Kang and Wong, 2018) via random projections, to estimating spectral densities of large matrices (Adams et al., 2018), to generating sketches of large-scale data (Pratap and Kulkarni, 2021; Pratap et al., 2021), and to sampling from sparse data (Li et al., 2008).

Figure 1 shows the common theme in the above applications. Instead of explicitly computing the desired quantity (e.g. inner products, matrix trace), a random variable from a chosen probability distribution is generated, and an estimator (a function of the data and the random variable) of the desired quantity is found. These applications further make use of information about the parameters in conjunction with maximum likelihood estimators (MLE) or a control variate estimator (CVE). This reduces the variance of these estimates, resulting in less error on average.

Information about parameters can come from the structure of data, e.g. observations in data might be normalized to a length of 1 or centered with zero mean, and this information can be used to get better estimates (Li et al., 2006). Marginal information could also be cheaply generated on the fly, where estimates with lower variance outweigh the cost of computing marginal information (Kang, 2021).

When both MLEs and CVEs are used for the same purpose, e.g. feature hashing (Verma et al., 2022) or random projections (Kang, 2021), the asymptotic theoretical variance reductions are identical. This is unsurprising, as MLEs are strongly consistent and asymptotically efficient under mild conditions and achieves the Cramér Rao lower bound (Lehmann, 1999). When the number of observations (i.e. size of sketch) is small,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

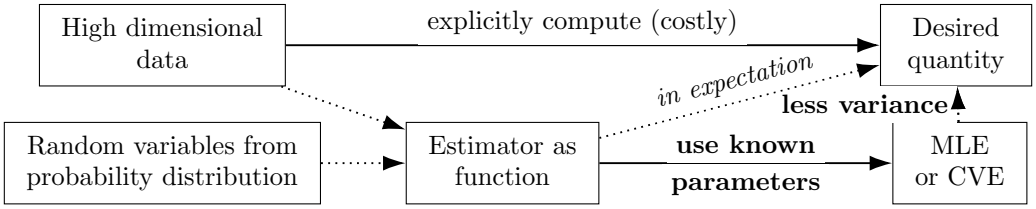

Figure 1: Idea behind estimation via sketching algorithms. **The goal is to use known parameters with MLE** *or* **CVE to get a reduced variance estimate**. We combine **MLE** *and* **CVE** in Algorithm 1 to get an estimator converging to the MLE faster with less numerical error than MLE/CVE alone, under mild conditions.

CVEs can outperform MLEs. *Yet there are two core problems behind using CVEs.*

Problem One: Replicating the CVE performance on same datasets give mixed results, leading to reproducibility issues. We give an explanation why this is so (and resolve this) in Sections 4 and 5.

Problem Two: Variance calculations with CVEs involve theoretical parameters for CV weights. But if CV weights rely on estimated data, then the theoretical variance is incorrect, hence it is tricky to analyze CVE performance. We propose Algorithm 1 with discussion in Appendix B.3 and B.4.

These problems disappear when the number of observations is large, but the purpose of using sketching algorithms is to reduce error (variance) with as few observations as possible. **Our paper aims to resolve these two issues and understand the performance of CVEs in sketching algorithms.**

We consider distributions from exponential families, previously known as Koopman-Darmois-Pitman families (Efron, 2022), and give examples across three common exponential families. Feature hashing (Weinberger et al., 2009; Verma et al., 2022) and estimating inner products (Indyk and Motwani, 1998; Li et al., 2006)) involve the Normal distribution. Frequency estimators (Cormode and Muthukrishnan, 2005; Pratap and Kulkarni, 2021) and angle estimation (Charikar, 2002; Kang and Wong, 2018) involve the multinomial distribution. Estimating a matrix trace (Hutchinson, 1989; Adams et al., 2018) involves the χ^2 distribution. We use geometric properties of exponential families to prove results in this paper. As exponential families may be unfamiliar to some readers, we describe an example of the zero-mean bivariate Normal in Appendix A to give intuition.

1.1 REVIEW OF EXPONENTIAL FAMILIES

Let $\mathcal{X} := \{\vec{x}_i\}_{i=1}^n$ comprising n observations $\vec{x}_i \in \mathbb{R}^D$. Let $p(\mathcal{X}|\vec{\nu})$ be a probability density function and $\vec{\nu} \in \mathbb{R}^p$ containing the parameters of the distribution. We write in exponential family form

$$p\left(\mathcal{X} \mid \vec{\nu}\right) \equiv \frac{\exp\left\{n\left(\sum_{j=1}^p \eta_j(\vec{\nu}) y_j(\mathcal{X})\right)\right\}}{\exp\left\{n\psi(\vec{\eta}(\vec{\nu}))\right\}} g\left(\mathcal{X}\right) = \frac{\exp\left\{n\vec{\eta}^T\vec{y}\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g\left(\mathcal{X}\right). \quad (1)$$

The vector $\vec{\eta} \in \mathbb{R}^p$ gives the canonical parameters of the distribution, and is (usually) not $\vec{\nu}$. There is a 1-1 map between $\vec{\eta}$ and $\vec{\nu}$, where each η_j can be expressed as a function of $\nu_1, \ldots, \nu_p$, i.e. $\eta_j = f_j(\nu_1, \ldots, \nu_p)$ and equivalently $\nu_j = g_j(\eta_1, \ldots, \eta_p)$. We use $\eta_j(\vec{\nu})$ in Equation (1) to denote η_j as a function of $\vec{\nu}$, but drop the parenthesis containing $\vec{\nu}$ for ease of notation. The components of $\vec{y}$ are sufficient statistics in terms of the data $\mathcal{X}$ corresponding to components of $\vec{\eta}$, where the y_is are linearly independent, and we also drop the parentheses containing $\{\vec{x}_i\}_{i=1}^n$ in y_j for ease of notation. $\psi(\vec{\eta})$ is the scaling factor which allows the probability density to integrate to 1. $g\left(\mathcal{X}\right)$ is independent of $\vec{\eta}$. Each (η_i, y_i) pair is unique up to an invertible linear transformation of the sufficient statistics. In particular, let $\mathbf{A}_{p\times p}$ be invertible, and define $\widetilde{y} = \mathbf{A}^{-1}\vec{y}$, $\widetilde{\eta} = \mathbf{A}^T\vec{\eta}$, and $\widetilde{\psi}(\widetilde{\eta}) = \psi\left(\mathbf{A}^{-T}\widetilde{\eta}\right)$. Then

$$p(\mathcal{X} \mid \vec{\nu}) = \frac{\exp\left\{n\left(\mathbf{A}^T\vec{\eta}\right)^T\left(\mathbf{A}^{-1}\vec{y}\right)\right\}}{\exp\left\{n\psi\left(\vec{\eta}\right)\right\}} g\left(\mathcal{X}\right) = \frac{\exp\left\{n\widetilde{\eta}^T\widetilde{y}\right\}}{\exp\left\{n\widetilde{\psi}\left(\widetilde{\eta}\right)\right\}} g\left(\mathcal{X}\right). \quad (2)$$

1.2 OUR CONTRIBUTIONS AND PAPER ORGANIZATION

Suppose data comes from a probability distribution belonging to an exponential family, and we want an estimate of unknown parameters from the distribution, given that some parameters are known. Under this framework, the **main goal** of this paper is to *demonstrate an equivalence between MLE and CVE* (**Contribution One**), leading to two important results. a)

Reproducibility between algorithms using MLE / CVE (**Contribution Two**), and b) *finding MLEs when the CV weights are known* (**Contribution Three**).

Contribution One: We formally show conditions for equivalence of asymptotic variance estimates via CVE and MLE and state an informal version of Theorem 4 which appears in Section 3.

Theorem 1. *(Informal Theorem 4)*
Suppose our observations come from an exponential family, with parameters divided into a set $\mathcal{A}$ to be estimated, and $\mathcal{B}$ which is known, and each $\mu_i \equiv \frac{\partial \psi(\vec{\eta})}{\partial \eta_i} = \mathbb{E}(y_i)$ is equal to ν_i. Then the asymptotic variance reduction of the MLE in estimating parameters in $\mathcal{A}$ is equal to the variance reduction given by a CVE where components in $\mathcal{B}$ are used as control variates.

We give a high level overview on how to prove Theorem 4 in Section 3, and put the main proof in Appendix B.

While Theorem 4 sounds like a known theorem as no other estimator can do better asymptotically than the MLE, there is no formal proof of CVE being equivalent to MLE to the best of our knowledge. For example: Glynn and Szechtman (2002) looks at an equivalence of CVE to non-parametric MLE, while we look at equivalence under exponential families. Reference books on exponential families (McLachlan and Krishnan, 2007; Efron, 2022) do not give this equivalence. The closest references to hint at an equivalence come from information geometry (Amari, 1987; Amari and Nagaoka, 2000), where to show asymptotic equivalence between the CVE and the MLE, then the optimal CV weights *must minimize the variance within the tangent space of the score functions and making the ancillary submanifold orthogonal within the local linearization around the model manifold*. We remark that these references only mention the condition for equivalence, but do not mention CVE at all.

We therefore prove Theorem 4 using results from exponential families rather than information geometry, in order to give intuition behind our work and make it accessible to more practitioners.

Contribution Two: Given Theorem 4, Theorem 6 shows that the CVE can be written as fixed point equations that are equivalent to the likelihood stationary equations. Under uniqueness and local convergence conditions in Corollary 2 and Theorem 7, Algorithm 1 converges to the MLE from sufficiently close initial values. We find that Algorithm 1 is empirically faster and numerically stable compared to several root-finding methods for the MLE with feature hashing (Verma et al., 2022) and random projection (Li et al., 2006; Kang, 2021) and can also resolve the reproducibility issue of CVE.

Contribution Three: We describe two heuristics that leads to efficiently finding the MLE via CVE weights (via Algorithm 1). We describe these heuristics in Section 6 and Appendix D. More specifically, we can get expressions for the MLE that are tedious to compute such as Hutchinson's Trace Estimator.

Section 2 of our paper reviews MLEs and CVEs. Section 3 describes our main results, and Section 4 discusses the implications of our results. Section 5 gives an example of how our work can be used to ensure reproducibility in evaluating algorithms using MLE/CVE. Section 6 demonstrates an example of how our results can be used to find a better (and more efficient) estimate (MLE) from CVE weights. Finally, we conclude in work in Section 7.

2 BACKGROUND INFORMATION AND SETTING UP THE PROBLEM

We give two theorems related to exponential families which are used in the main proof of Theorem 4.

Theorem 2. *Given n observations, the mean vector $\vec{\mu} \equiv \mathbb{E}(\vec{y})$ and the variance-covariance matrix $\mathbf{V}_n \equiv \mathrm{Cov}(\vec{y})$ for $\vec{y}$ in the exponential family is given by $\vec{\mu} = \left(\frac{\partial \psi(\vec{\eta})}{\partial \eta_1}, \ldots, \frac{\partial \psi(\vec{\eta})}{\partial \eta_p}\right)$ and $n\mathbf{V}_n = \mathbf{V}$ where*

$$\mathbf{V} = \begin{pmatrix} \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_1^2} & \cdots & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_1 \partial \eta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_p \eta_1} & \cdots & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_p^2} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & \cdots & \frac{\partial \mu_1}{\partial \eta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_p}{\partial \eta_1} & \cdots & \frac{\partial \mu_p}{\partial \eta_p} \end{pmatrix}.$$

Proof. A proof can be found in Efron (2022) and in the appendix on Page 16. □

Theorem 3. *(Efron, 1978, 2022) In exponential families, the corresponding functions $\vec{\eta}$ and $\vec{\mu}$ have 1-1 mappings, given by the differential relationship $d\mu = \mathbf{V} d\eta$, and $d\eta = \mathbf{V}^{-1} d\mu$.*

2.1 REVIEW OF MAXIMUM LIKELIHOOD ESTIMATION

In exponential families, the log-likelihood with respect to $\vec{\eta}$ is $l(\mathcal{X}|\vec{\eta}) = n\left(\vec{\eta}^T \vec{y} - \psi(\vec{\eta})\right) + \log(g(\mathcal{X}))$, and the score function with respect to $\vec{\eta}$ is given by $l'(\mathcal{X}|\vec{\eta}) = \frac{\partial l(\mathcal{X}|\vec{\eta})}{\partial \vec{\eta}} = n(\vec{y} - \vec{\mu})$ by applying Theorem 2. The MLE for $\vec{\eta}$ is achieved with $\hat{\mu}_i = y_i$ by setting $l'(\mathcal{X}|\vec{\eta})$ to be 0. The Fisher Information with respect to $\vec{\eta}$ is $i_{\vec{\eta}} = \mathbb{E}\left(l'(\mathcal{X}|\vec{\eta}) l'(\mathcal{X}|\vec{\eta})^T\right) = nV_n$. The distribution of the MLE of $\vec{\eta}$ converges to a multivariate Normal $\mathcal{N}(\vec{\eta}, i_{\vec{\eta}}^{-1})$ (Lehmann, 1999). The MLE of $\vec{\nu}$ is found by computing the score function with respect to $\vec{\nu}$

and equating it to zero. Equivalently, if the MLEs $\hat{\eta}_i$ have been computed under the exponential family framework, then the MLE of $\hat{\nu}_i = g_i(\hat{\eta}_1, \ldots, \hat{\eta}_p)$.

The MLE of $\hat{\nu}$ similarly converges to $\mathcal{N}(\vec{\nu}, i_{\vec{\nu}}^{-1})$, where $i_{\vec{\nu}}$ is the Fisher Information of $\vec{\nu}$. Without loss of generality, let $\nu_{\mathrm{E}} := \{\nu_i\}_{i=1}^{t}$ be the set of parameters to be estimated, and $\nu_{\mathrm{K}} := \{\nu_j\}_{j=t+1}^{p}$ be the set of parameters known. Instead of each η_i as a function $f_i(\nu_1, \ldots, \nu_p)$, we restrict η_i to ν_{E}, i.e. $\eta_i = \tilde{f}_i(\nu_1, \ldots, \nu_t)$, and write this as $\vec{\eta}(\nu_{\mathrm{E}})$. With some abuse of notation, the log-likelihood is restricted to $\nu_{\mathrm{E}} \in \mathbb{R}^t$, i.e. $l\left(\mathcal{X} \mid \vec{\eta}(\nu_{\mathrm{E}})\right) = n\left(\vec{\eta}(\nu_{\mathrm{E}})^T \vec{y} - \psi\left(\vec{\eta}(\nu_{\mathrm{E}})\right)\right) + \log(g(\mathcal{X}))$, and the score function with respect to ν_{E} is

$$\begin{aligned}\frac{\partial\, l(\mathcal{X} |\, \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}} &= n\left(\frac{\partial \vec{\eta}(\nu_{\mathrm{E}})}{\partial \nu_{\mathrm{E}}}\right)^T \vec{y} - n\left(\frac{\partial \psi\left(\vec{\eta}(\nu_{\mathrm{E}})\right)}{\partial \vec{\eta}(\nu_{\mathrm{E}})}\right)^T \frac{\partial \vec{\eta}(\nu_{\mathrm{E}})}{\partial \nu_{\mathrm{E}}} \\ &= n\left(\frac{\partial \vec{\eta}(\nu_{\mathrm{E}})}{\partial \nu_{\mathrm{E}}}\right)^T (\vec{y} - \vec{\mu})\,. \end{aligned} \tag{3}$$

We defer computing the Fisher Information with respect to ν_{E} to the proof of Theorem 4.

2.2 REVIEW OF CONTROL VARIATES

Suppose we want to estimate the mean of a random variable X, given by $\mathbb{E}(X)$. Further suppose we know the random process that generated X, and can use the same process to generate random variables $Y_1, \ldots, Y_k$, where $\mathbb{E}(Y_1), \ldots, \mathbb{E}(Y_k)$ are known. Then $Z = X + \sum_{i=1}^{k} c_i(Y_i - \mathbb{E}(Y_i))$ is a CVE for X. The Y_is are called control variates (CV). $\mathbb{E}(Z) = \mathbb{E}(X)$, but the variance of Z is given by $\mathrm{Var}(Z) = \mathrm{Var}(X) + \sum_{i=1}^{k} c_i^2 \mathrm{Var}(Y_i) + 2\sum_{i=1}^{k} c_i \mathrm{Cov}(X, Y_i) + 2\sum_{i>j}^{k} c_i c_j \mathrm{Cov}(Y_i, Y_j)$. Let the matrix $\mathbf{W}_{ij} = \mathrm{Cov}(Y_i, Y_j)$, $1 \leq i, j, \leq k$, noting that $\mathrm{Cov}(Y_i, Y_i) \equiv \mathrm{Var}(Y_i)$, and let $\vec{d} = (\mathrm{Cov}(X, Y_1), \ldots, \mathrm{Cov}(X, Y_k))^T$. Then the optimal values of $\vec{c}$ minimizing $\mathrm{Var}(Z)$ are given by $\mathbf{W}\vec{c} = -\vec{d}$ and are called the CV corrections. The optimal variance given by the CVE is

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}(X) + \vec{c}^T \mathbf{W} \vec{c} + 2\vec{c}^T \vec{d} \\ &= \mathrm{Var}(X) + \vec{c}^T \vec{d} = \mathrm{Var}(X) - \vec{d}^T \mathbf{W}^{-1} \vec{d}. \end{aligned} \tag{4}$$

From Equation (4), $\mathrm{Var}(Z) \leq \mathrm{Var}(X)$, with equality only when $\vec{d} = \vec{0}$ since the covariance matrix $\mathbf{W}$ is positive semi-definite and invertible, hence $\mathbf{W}^{-1}$ is positive semi-definite as well. Suppose $\nu_{\mathrm{E}} = \{\nu_i\}_{i=1}^{t}$ and we want an estimate of some linear combination of $X = \sum_{i=1}^{t} \alpha_i \nu_i, \alpha_i \in \mathbb{R}$, with $\nu_{\mathrm{K}} = \{\nu_j\}_{j=t+1}^{p}$. We make the following assumptions.

Assumption One: A CV leading to the most variance reduction when parameters are linear must necessarily come from the sufficient statistics y_j as sufficient statistics yield the most information about the data. We look at linear combinations of y_js, rather than any arbitrary functions of the y_js, due to the form of the exponential family. We stress that we are not focusing on coming up with zero-variance CVEs along the lines of Oates et al. (2017); South et al. (2023), but more of relating the MLE to the CVE.

Assumption Two: Given $(p - t)$ independent known parameters $\nu_{t+1}, \ldots, \nu_p$, there ought to exist $(p-t)$ CV with known means. We describe how to find them in (the proof of) Corollary 1.

3 EQUIVALENCE OF MLE AND CV

We state our main theorem, and give the high level overview on our proof strategy.

Theorem 4. *Let observations $\mathcal{X} := \{\vec{x}_i\}_{i=1}^{n}$ come from an exponential family which is regular and minimal, with ψ twice continuously differentiable and $\mathbf{V} = \nabla^2 \psi(\vec{\eta}) \succ 0$ at the parameter value under consideration. Let $\nu_{\mathrm{E}} := \{\nu_i\}_{i=1}^{t}$ be the set of parameters to be estimated, and $\nu_{\mathrm{K}} := \{\nu_j\}_{j=t+1}^{p}$ be the set of parameters known. Suppose every $\mu_i \equiv \frac{\partial \psi(\vec{\eta})}{\partial \eta_i} = \mathbb{E}(y_i)$ is equal to ν_i. The asymptotic variance reduction given by the MLE of a linear combination of ν_{E} is equal to the variance reduction given by a CVE where components in $\vec{y}$ are used as control variates.*

The tricky part of proving Theorem 4 is to show that the **asymptotic variance** of the MLE is the same as the **variance given by the CVE**, by "standardizing" the (very different!) terminology in both fields to exponential family notation.

In brief, we compute the Fisher Information with respect to ν_{E} (from Equation 3) which involves the matrix of partial derivatives $\left[\frac{\partial \eta_i}{\partial \mu_j}\right]_{i,j}, 1 \leq i, j \leq t$ in order to get an expression for the **asymptotic variance of the MLE**. We compute the expression for the **variance of the CVE** which involves the matrix of partial derivatives $\left[\frac{\partial \mu_i}{\partial \eta_j}\right]_{i,j}, 1 \leq i, j \leq t$. We then apply Lemma 1 (in Appendix B) to show how the two matrices of partial derivatives are related, and finally prove the equivalence. Figure 2 gives the gist of our approach, and a full and complete proof is on Page 17 in Appendix B.2.

Corollary 1. *Theorem 4 also holds when $\vec{\mu} = \mathbf{A}\vec{\nu}$.*

Proof. This follows since we can make the transformation in Equation (2) with $\widetilde{\eta} = \mathbf{A}\vec{\eta}$ and $\widetilde{y} = \mathbf{A}^{-1}\vec{y}$ for $\mathbf{A}$ an invertible matrix of appropriate size, before applying Theorem 4. A more detailed proof is on Page 19 in the appendix. □

Theorem 4 shows that the MLE and the optimal CVE have the same asymptotic variance reduction. We show

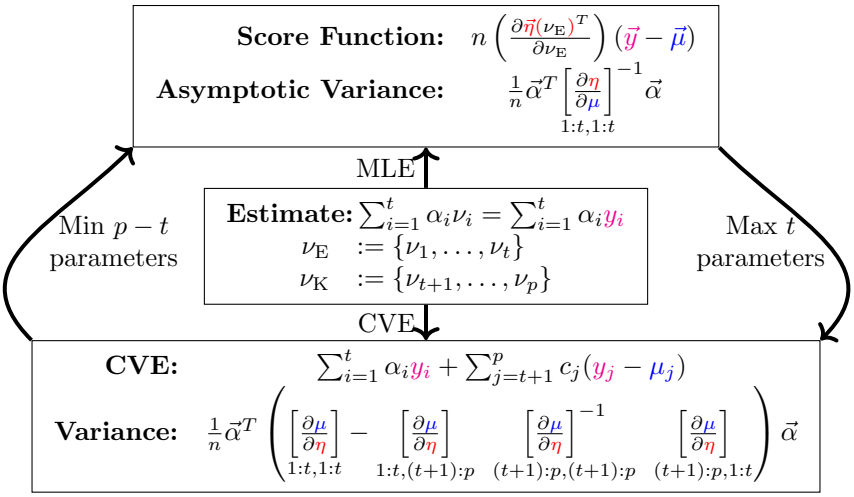

Figure 2: The exponential family $\frac{\exp\left\{n\vec{\eta}(\nu_{\mathrm{E}},\nu_{\mathrm{K}})^T\vec{y}\right\}}{\exp\left\{n\psi(\vec{\eta}(\nu_{\mathrm{E}},\nu_{\mathrm{K}}))\right\}}g(\mathcal{X})$ for $\mathbf{V}^{MLE}=\mathbf{V}^{CVE}$.

in Appendix B that the likelihood score equations for ν_{E} can be written in the same form as the optimal CV coefficients. In particular, for each $1\le i\le t$, the MLE satisfies a fixed point equation $\nu_i = f_i(\nu_1,\ldots,\nu_t) = y_i+\sum_{j=t+1}^p\hat{c}_{ij}(\nu_1,\ldots,\nu_t)(y_j-\mu_j)$ where $\hat{c}_{ij}$s are the optimal CV coefficients and $\mu_j=\nu_j$ is known for $j>t$. Hence, we get a system of fixed point equations whose solutions are the stationary points of the MLE, which we describe in Algorithm 1.

4 DISCUSSION

For $\nu_{\mathrm{E}} := \{\nu_i\}_{i=1}^t$, $\nu_{\mathrm{K}} := \{\nu_j\}_{j=t+1}^p$, the method of MLE for estimating ν_{E} maximizes the score function with respect to ν_{E} over t terms. The asymptotic variance of the MLE of ν_{E} is found via the inverse of the Fisher Information. Statistical theory states that an unbiased MLE asymptotically achieves the Cramér Rao lower bound, implying no other estimator can do better (Lehmann, 1999). However, finding the optimal ν_{E} usually involves numerical root-finding methods. On the other hand, the method of CV for estimating ν_{E} minimizes the variance of the CVE, and is minimized over $(p-t)$ terms. We make three observations, with Figure 2 showing a summary of the equivalence.

Observation 1: There is a duality between the variance estimates given by the MLE (maximizing t terms in $\frac{\partial\eta}{\partial\mu}$, by *knowing the likelihood function*) and by the CVE (minimizing $p-t$ terms in $\frac{\partial\mu}{\partial\eta}$, *by computing relevant second moments*). In cases where we know fewer parameters $p-t$, then using CVE is desirable. Otherwise, using the MLE is desirable.

Observation 2: Solving for the MLE requires the score function with respect to $\vec{\nu}$ without exponential families. Under the exponential family framework, we must know the map from $\vec{\eta}$ to $\vec{\nu}$ to compute partial derivatives of the form $\frac{\partial\eta_i}{\partial\nu_j}$ in order to find the MLE as in Equation (3). For the CVE, we require partial derivatives of the form $\frac{\partial\mu_i}{\partial\eta_j}$ in order to find the optimal control variate coefficients $\hat{c}_i$ leading to the CV estimate. Using one method over the other depends on what information we have.

Observation 3: In most cases, the covariances in the optimal CV corrections $\hat{c}_i$ is a function of ν_{E}. In Equation (19), solving for $\hat{c}_i$ requires knowing the respective $\mathrm{Cov}\left(y_i,y_s\right)\equiv\mathbb{E}\left(y_iy_s\right)-\mathbb{E}\left(y_i\right)\mathbb{E}\left(y_s\right), t+1\le s\le p$, yet the terms y_i are what we wanted to estimate and are unknown. This invalidates the theoretical variance calculations of the CVE if substitution of the terms y_i are used. However, Algorithm 1 implies that the theoretical variance **is** the MLE theoretical variance.

Various papers treat the CV correction $\hat{c}_i$ differently: e.g. Adams et al. (2018) computes the empirical covariance and variance in a CVE for Hutchinson's Estimator, while Kang (2021) uses initial estimates of ν_{E} to substitute into respective $\hat{c}_i$s, as well as a constructing a multivariate Normal distribution where terms involving ν_{E} cancel out in the CV correction for random projections. Our results imply that treating the CVE as a fixed point algorithm in our experiments result in a better and *theoretically correct* estimator, in the sense that the variance of the estimator after convergence is equal to the variance of the MLE at that specific number of observations (or sketch size) k. We further suggest when conditions in Theorem 4 or Corollary 1 hold, (`CV-FP`) is empirically faster and numerically stable compared to root finding algorithms such as Newton Raphson and the Secant method, *and mitigates the*

Algorithm 1: Fixed Point Algorithm via Control Variates (`CV-FP`)

Preconditions: Want to estimate some linear combination of $\nu_{\mathrm{E}} = \{\nu_1, \ldots, \nu_t\}$ and know $\nu_{\mathrm{K}} = \{\nu_{t+1}, \ldots, \nu_p\}$

Require: t control variate expressions of the form $f_i(\nu_1, \ldots, \nu_t) = y_i + \sum_{j=t+1}^{p} \hat{c}_{ij}(\nu_1, \ldots, \nu_t)(y_j - \mu_j)$ where $\hat{c}_{ij}(\nu_1, \ldots, \nu_t)$ can be evaluated

Result: A fixed point of the CVE, which is the MLE of ν_{E} under Corollary 2

Get initial estimates of $\hat{\nu}_1^{(1)} = y_1, \ldots, \hat{\nu}_t^{(1)} = y_t$ and initialize $n = 1$;

repeat

$\hat{\nu}_1^{(n+1)} = f_1(\hat{\nu}_1^{(n)}, \hat{\nu}_2^{(n)}, \ldots, \hat{\nu}_t^{(n)})$;

$\hat{\nu}_2^{(n+1)} = f_2(\hat{\nu}_1^{(n+1)}, \hat{\nu}_2^{(n)}, \ldots, \hat{\nu}_t^{(n)})$;

$\ldots$;

$\hat{\nu}_t^{(n+1)} = f_t(\hat{\nu}_1^{(n+1)}, \hat{\nu}_2^{(n+1)}, \ldots, \hat{\nu}_t^{(n)})$;

$n = n + 1$;

until *all* $|\hat{\nu}_s^{(n)} - \hat{\nu}_s^{(n-1)}| \leq \epsilon$, *for* $1 \leq s \leq t$;

Return $(\hat{\nu}_1^{(n)}, \ldots, \hat{\nu}_t^{(n)})$;

reproducibility issues mentioned in our introduction.

While Newton Raphson has quadratic convergence, the derivative has to be computed at each update without guarantee of convergence at bad starting points even if the initial estimate of y_i was used. The Secant method has superlinear convergence, though it requires two initial starting observations, which have to be near the root, although they can be some $y_i \pm \epsilon$, or the upper and lower bounds of what y_i could be. Empirically, our experiments in Section 5 for the bivariate Normal distribution show a better convergence to the root with `CV-FP` compared to Newton Raphson for the MLE, and gives a reasonable explanation for the differing performance of CVEs compared to MLEs.

5 EXPERIMENTS

We demonstrate the practical applications of this equivalence on two algorithms: feature hashing and random projection with respect to *reproducibility*. We briefly discuss the equivalence between CVE and MLE for feature hashing in the main paper and defer to Appendix C for a detailed discussion on both algorithms.

Feature hashing is a technique to quickly estimate inner products between vectors $\vec{x}_i, \vec{x}_j$, with applications in similarity search, duplicate detection (Nauman and Herschel, 2022), etc. It compresses the vectors to low-dimensional sketches and estimates the inner product from them. Suppose $\vec{v}_i, \vec{v}_j \in \mathbb{R}^k$ are the feature hashing sketch of the vectors $\vec{x}_i, \vec{x}_j \in \mathbb{R}^p$ where $k \ll d$. Then $\langle \vec{v}_i, \vec{v}_j \rangle$ gives an unbiased estimate of $\langle \vec{x}_i, \vec{x}_j \rangle$ (Weinberger et al., 2009). We apply Algorithm 1 (`CV-FP`) to Verma et al. (2022) to estimate $\langle \vec{x}_i, \vec{x}_j \rangle$, getting the update step of

$$f_{n+1} = \sum_{s=1}^{k} v_{is} v_{js} - \frac{f_n \|\vec{x}_j\|^2 \left(\sum_{s=1}^{k} v_{is}^2 - \|\vec{x}_i\|^2 \right)}{f_n^2 + \|\vec{x}_i\|^2 \|\vec{x}_j\|^2} - \frac{f_n \left(\|\vec{x}_i\|^2 \left(\sum_{s=1}^{k} v_{js}^2 - \|\vec{x}_j\|^2 \right) \right)}{f_n^2 + \|\vec{x}_i\|^2 \|\vec{x}_j\|^2} \tag{5}$$

where we set $f_1 = \sum_{s=1}^{k} v_{is} v_{js}$. Further details are in Appendix C.

Hardware Description: Experiments were done on a machine having the following configuration: CPU: Intel(R) Core(TM) i7-8750H CPU @ 2.21GHz x 6; Memory: 16 GB; OS: Windows 10.

We run simulations on randomly generated vector pairs $\vec{x}_1, \vec{x}_2 \in \mathbb{R}^{10,000}$ with ratios $r \in \{0.1, 0.5, 1, 2, 10\}$ where $\|\vec{x}_1\|^2 = r\|\vec{x}_2\|^2$, and angles $\theta \in \{\frac{\pi}{12}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \frac{11\pi}{12}\}$. Each specific ratio and angle for a vector pair takes ≈ 2 hours to run, and data generated can be up to 150MB.

We generate feature hashing sketches, denoted as $\vec{v}_i$ and $\vec{v}_j$, for vector pairs $\vec{x}_i$ and $\vec{x}_j$ respectively for sketch size (k) where $1 \leq k \leq 100$. We compute $\sum_{s=1}^{k} v_{is} v_{js}$ and refer to it as the baseline estimate. Subsequently, we compute estimates using: a) MLE (Verma et al., 2022) via Newton Raphson (`MLE-NR`), b) MLE via the Secant method (`MLE-Secant`), c) CV using $\sum_{s=1}^{k} v_{is} v_{js} \approx \langle \vec{x}_i, \vec{x}_j \rangle$ in $\hat{c}_i$ (`CV-Init`), d) CV where the empirical covariance and variance are computed for $\hat{c}_i$ (`CV-Emp`), and e) Algorithm 1 (`CV-FP`). We repeat this for $10,000$ iterations, and record the number of updates a) b) and e) take until convergence.

Figure 3 summarizes the mean square error (MSE) across $10,000$ iterations for various sketch sizes (k), considering $\|\vec{x}_1\|^2 = \|\vec{x}_2\|^2$ and $\theta = \frac{\pi}{12}$. A lower MSE indicates better performance. We display $\theta = \frac{\pi}{12}$ since similar vectors are usually of interest with more plots in Appendix C. From Figure 3, it is clear that the MSE of all five methods is lower than that of the baseline feature hashing estimate. `CV-FP` and `MLE-Secant` show similar best performances compared to others. Unsurprisingly, out of the control variate implementions, `CV-Init` is the worse, followed by `CV-Emp`, with `CV-FP` being the best, since the theoretical variance calculations are invalidated with empirical values, and only `CV-FP` achieves the MLE variance reduction.

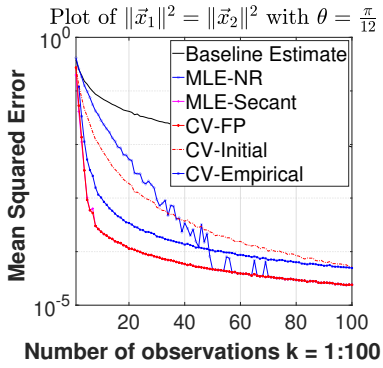

Figure 3: MSE plot when $\|\vec{x}_1\|^2 = \|\vec{x}_2\|^2$ with $\theta = \frac{\pi}{12}$.

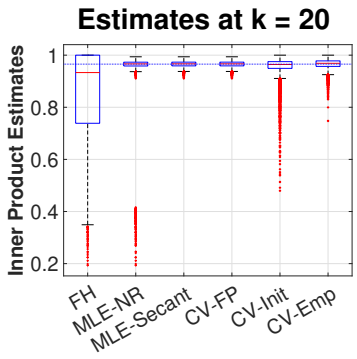

Figure 4: Boxplots of estimated inner products. Blue horizontal line denotes true inner product.

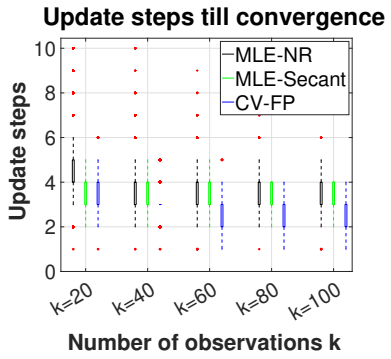

Figure 5: Boxplots of update steps till convergence at $k = \{20, 40, 60, 80, 100\}$

At smaller values of k, `MLE-NR` has the worst performance, and after a threshold ($k \approx 70$), its performance converges to the performance of `MLE-Secant` and `CV-FP`. Instead of error bars, we give boxplots in Figure 4 for the respective estimates for these algorithms at sketch size $k = 20$. We can see the poor performance of `MLE-NR` is mostly due to a higher number of outliers, since the interquartile range is comparable to other methods. Figure 5 gives boxplots of update steps till convergence for `MLE-NR`, `MLE-Secant` and `CV-FP`. `CV-FP` generally takes fewer number of steps compared to the other two baselines, indicating the faster empirical convergence of `CV-FP`. Additional plots across all angles and norms, discussion of converging to the wrong estimate and convergence rates of a), b), and e) are in Appendix B.4 and C.

Implication of our experiments on reproducibility: Most papers involving MLEs and CVEs focus on theoretical work (proving variance bounds) and experiments (comparing against benchmarked algorithms), but do not explicitly mention how respective estimates from the CVE and MLE are found, as the assumption is that a user can implement them on their own. We suspect this is why there are mixed empirical results when CVE and/or MLE are used.

Our plots show there is a great difference in the methods used: from numerical error (Figure 3) to speed of convergence (Figure 5), if a suboptimal implementation of `Newton Raphson` *is used for the MLE, or computing* $\hat{c}$ *via* `CV-Init` *(or* `CV-Emp`*) via the CVE.* This gives contradictory results based on the type of experiments.

For example, identifying vectors with high similarity is a goal of similarity search, where the angle θ between the two vectors is small. A difference in root finding methods used can lead to different performance, e.g. `Newton Raphson` has poor performance with small k (Figure 4), but that does not mean the MLE is bad. Equivalently, researchers may use the median (which does not take into account outliers) to show better results (boxplots in Figure 4 show `MLE-NR` has similar (good) performance with `MLE-Secant` and `CV-FP` ignoring outliers); but is not realistic in a practical scenario where outliers are not known in advance. Moreover, if sole error bars are used (based on computing standard deviations of the estimates), this may lead to a conclusion that an empirical MLE has "higher error", even though the outliers increase standard deviations (Figure 4) and can be misleading. Figure 4 shows there are more outliers in one direction for Newton Raphson. Using error bars implicitly assumes "equal variation above and below the mean", and masks what is happening in practice.

Algorithm 1 (`CV-FP`) can mitigate these issues.

6 FINDING THE MLE VIA CVE WEIGHTS

Two heuristics arising from the construction of `CV-FP` can be used to find the MLE, even if the conditions in Theorem 4 or Corollary 1 are not satisfied, by numerically solving the fixed point iteration, or solving it analytically.

While we describe an application with respect to sketching algorithms, we believe our technique can be applied in different fields (e.g. Bayesian updates via finite mixture models) where once we find the CVE weights, we can repeatedly iterate over them, or find a closed form solution which gives the MLE.

We give an example for Hutchinson's estimator (Hutchinson, 1989), which is used to estimate the trace of a matrix $\mathbf{M}$; but has found use in many applications, e.g. estimating the log determinant of a matrix (Cortinovis and Kressner, 2021) (using the fact that $\log(\det(\mathbf{M})) = \mathrm{tr}\,(\log(\mathbf{M}))$), counting number of triangles in graphs (using the fact that number of triangles in

graph $G = \frac{1}{6}\text{tr}\left(\mathbf{M}^3\right)$, where $\mathbf{M}$ is the adjacency matrix of G) (Avron, 2010). Improved versions of Hutchinson's Estimator, such as Hutch++ (Meyer et al., 2021) also involve estimating the trace of a (different) matrix, and reducing the variance of the trace estimation in these cases is desirable.

Let $\vec{r}_i \in \mathbb{R}^d$ be distributed $\mathcal{N}(\vec{0}_d, \mathbf{I}_d)$. Let $y_i = \vec{r}_i^T\mathbf{M}\vec{r}_i \in \mathbb{R}$ for $1 \leq i \leq k$. For a symmetric $\mathbf{M}_{d\times d}$ matrix, the estimator $Y = \sum_{i=1}^k \frac{y_i}{k} = \sum_{i=1}^k \frac{\vec{r}_i^T\mathbf{M}\vec{r}_i}{k}$ estimates the trace of $\mathbf{M}$, with $\text{Var}\left(Y\right) = \frac{2\text{tr}\left(\mathbf{M}^2\right)}{k} = \frac{2\|\mathbf{M}\|_F^2}{k}$, where $\|\mathbf{M}\|_F^2$ is the Frobenius norm, defined by $\|\mathbf{M}\|_F^2 = \sum_{i,j} m_{ij}^2$. Adams et al. (2018) gives the following CVE for $\text{tr}\left(\mathbf{M}\right)$ when there is a known diagonal matrix $\mathbf{B}$,

$$Z = \sum_{i=1}^k \frac{\vec{r}_i^T\mathbf{M}\vec{r}_i}{k} + c\left(\sum_{i=1}^k \frac{\vec{r}_i^T\mathbf{B}\vec{r}_i}{k} - \text{tr}\left(\mathbf{B}\right)\right),$$

and Lemma 4.1 in Adams et al. (2018) gives the optimal $\hat{c}$ to be

$$\hat{c} = -\frac{\text{Cov}\left(\vec{r}^T\mathbf{M}\vec{r}, \vec{r}^T\mathbf{B}\vec{r}\right)}{\text{Var}\left(\vec{r}^T\mathbf{B}\vec{r}\right)} = -\frac{\text{tr}\left(\mathbf{MB}\right)}{\text{tr}\left(\mathbf{B}^2\right)},$$

with a variance reduction of $\frac{2\text{tr}(\mathbf{MB})^2}{k\text{tr}(\mathbf{B}^2)}$. The empirical covariance $\text{Cov}\left(\vec{r}^T\mathbf{M}\vec{r}, \vec{r}^T\mathbf{B}\vec{r}\right)$ and variance $\text{Var}\left(\vec{r}^T\mathbf{B}\vec{r}\right)$ may be computed from the observed data, since $\text{tr}\left(\mathbf{MB}\right)$ is not known (so if empirical covariances and variances are used, the variance reduction is not theoretically correct).

Given the work in Adams et al. (2018), a natural question would be to ask if a MLE could be found, which might yield a lower variance reduction. However, Y is distributed as a linear combination of χ^2 variables, and finding an MLE is non-trivial.

Our first heuristic is to *find linearly independent sufficient statistics* to construct the CVE and find optimal CV corrections $\hat{c}_i$. We do so for each diagonal term $\mathbf{M}_{tt}$ in $\text{tr}\left(\mathbf{M}\right)$. The new CVE is given by $\mathbb{E}\left(\vec{r}^T\mathbf{M}\vec{r} + \sum_{i=1}^d c_i\left(\sum_{s=1}^d r_s\left(\mathbf{B}\vec{r}\right)_s - b_{ii}\right)\right)$. Detailed derivations for this heuristic is described in the proof of Theorem 8 in Appendix D.

The second heuristic is to *treat the CVE as a fixed point iteration (and perhaps find a closed form solution)*. In the case of Hutchinson's estimator, there exists a closed form solution, leading to the estimator $\widehat{\text{tr}\left(\mathbf{M}\right)} = \sum_{s=1}^d \frac{b_{ss}r_s(\mathbf{M}\vec{r})_s}{r_s(\mathbf{B}\vec{r})_s}$ with overall approximate variance reduction given by $\frac{2}{k}\sum_{i=1}^d m_{ii}^2$. We remark that even though the conditions for Theorem 4 and Corollary 1 are not satisfied, we can do standard variance analysis on any new estimator we find. Details are on Pages 37 and 38 in Appendix D.

The estimator we find under our two heuristics is Bekas' diagonal estimator (Bekas et al., 2007) (when Normal random variables are used). Hence the answer to the question proposed regarding the MLE would be: *The computations for the MLE could be bypassed via the CVE (and in fact, has the same form of Bekas et al. (2007)), and the MLE would be a biased estimator.* We remark that the original paper of Bekas et al. (2007) does not give any variance analysis of their estimator (or further intuition on constructing the estimator), though Baston and Nakatsukasa (2022) builds on their work to construct probability bounds, so our results in Appendix D can be thought of as a mini-technical report complementing both papers.

To summarize, even if conditions do not hold, numerically solving the fixed point iteration, or finding its closed form analytically can give new estimators, where its variance (and other properties) can be computed. With respect to our field in sketching algorithms, our heuristic can lead to re-examining past work on estimators to give alternate proofs for variance reduction (and using these variances in probability bounds), and future work on understanding new estimators along similar lines, e.g. applying our work to find better estimators for estimating Euclidean distances under composition of functions (Leroux and Rademacher, 2024), or for the Jaccard similarity under circulant permutations (Li and Li, 2022). In both these cases, computing expectations and covariances (hence CVE weights) are much easier than analytically deriving the MLE involved.

7 CONCLUSION, BROADER IMPACT, AND FUTURE WORK

We have shown a connection in exponential families between $\mathbf{V}_{ij} = \frac{\partial\mu_i}{\partial\eta_j}$ and $\mathbf{V}_{ij}^{-1} = \frac{\partial\eta_i}{\partial\mu_j}$ for CVE and MLE respectively, with `CV-FP` converging to the MLE, leading to two important results: *reproducibility* between algorithms using MLE / CVE, and *finding MLEs when the CV weights are known.* We briefly mention the limitations of our work, two broad impacts our work might have, and two directions for future research.

Limitations: Our work is limited to exponential families where each μ_i is a linear combination of ν_i (and vice versa), and our experiments only look at the bivariate Normal distribution. However, we expect the performance of `CV-FP` to generalize across distributions from exponential families under these conditions.

Unify Existing MLE/CVE: Showing that $\mathbf{V}^{\text{MLE}} = \mathbf{V}^{\text{CVE}}$ gives an algorithm (`CV-FP`) which is numerically stable and faster via experiments for the Normal distribution. `CV-FP` has the potential to lead to better reproducibility of estimates from existing MLE/CVE,

and ensure that evaluation of future estimators against benchmark MLE/CVE is fair when comparing against accuracy and speed.

Intuition For Existing / New Estimators: The heuristic of *finding linearly independent sufficient statistics* to construct a CVE, and *treating the CVE as a fixed point iteration* gives an estimator, even if conditions for Theorem 4 are not met. The variance reduction from the CVE also gives some intuition of how the initial data can affect variance reduction.

Future Work (Extending `CV-FP`): $\mathbf{V}^{\text{MLE}} = \mathbf{V}^{\text{CVE}}$ holds when νs are linear combinations of μs. There is analogous future work to show the relationship between $\mathbf{V}^{\text{MLE}}$ and $\mathbf{V}^{\text{CVE}}$ and convergence of `CV-FP` when this condition does not hold. We gave an example of the χ^2 distribution in Section 6 where we derived Bekas' estimator a different way, which hints that $\mathbf{V}^{\text{MLE}} = \mathbf{V}^{\text{CVE}}$ when νs are non linear combinations of μs.

Future Work (Convergence of `CV-FP`): Simulations in Appendix C.5 show `CV-FP` could either be linear or superlinear, with a small constant C. Future research could prove this convergence and give a bound on C for various probability distributions.

Acknowledgments: The authors would like to thank the many reviewers who contributed helpful suggestions, the Bucknell Physics and Astronomy Summer Research program, and Pamela Gorkin, Peter McNamara, Sanjay Dharmavaram, and Sara Stoudt for insightful conversations.

References

Adams, R. P., Pennington, J., Johnson, M. J., Smith, J., Ovadia, Y., Patton, B., and Saunderson, J. (2018). Estimating the Spectral Density of Large Implicit Matrices. *arXiv preprint arXiv:1802.03451*.

Amari, S.-i. (1987). *Differential geometrical theory of statistics*, volume 10. Hayward California.

Amari, S.-i. and Nagaoka, H. (2000). *Methods of information geometry*, volume 191. American Mathematical Soc.

Atkinson, K. E. (2008). *An Introduction to Numerical Analysis*. John Wiley & Sons.

Avron, H. (2010). Counting triangles in large graphs using randomized matrix trace estimation. In *Workshop on Large-scale Data Mining: Theory and Applications*, volume 10, page 9.

Baston, R. A. and Nakatsukasa, Y. (2022). Stochastic diagonal estimation: Probabilistic bounds and an improved algorithm. *arXiv preprint arXiv:2201.10684*.

Bekas, C., Kokiopoulou, E., and Saad, Y. (2007). An estimator for the diagonal of a matrix. *Applied numerical mathematics*, 57(11-12):1214–1229.

Charikar, M. S. (2002). Similarity Estimation Techniques From Rounding Algorithms. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pages 380–388.

Cormode, G. and Muthukrishnan, S. (2005). An Improved Data Stream Summary: the Count-Min Sketch and its Applications. *Journal of Algorithms*, 55(1):58–75.

Cortinovis, A. and Kressner, D. (2021). On randomized trace estimates for indefinite matrices with an application to determinants. *Foundations of Computational Mathematics*, pages 1–29.

Efron, B. (1978). The Geometry of Exponential Families. *The Annals of Statistics*, pages 362–376.

Efron, B. (2022). *Exponential Families in Theory and Practice*. Cambridge University Press.

Glynn, P. W. and Szechtman, R. (2002). Some New Perspectives on the Method of Control Variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000: Proceedings of a Conference held at Hong Kong Baptist University, Hong Kong SAR, China, November 27–December 1, 2000*, pages 27–49. Springer.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. JHU Press.

Hutchinson, M. F. (1989). A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076.

Indyk, P. and Motwani, R. (1998). Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, pages 604–613.

Kang, K. (2021). Correlations Between Random Projections and the Bivariate Normal. *Data Mining and Knowledge Discovery*, 35(4):1622–1653.

Kang, K. and Wong, W. P. (2018). Improving Sign Random Projections with Additional Information. In *International Conference on Machine Learning*, pages 2479–2487. PMLR.

Keener, R. W. (2010). *Theoretical Statistics: Topics for a Core Course*. Springer Science & Business Media.

Lavenberg, S. S. and Welch, P. D. (1981). A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations. *Management Science*, 27(3):322–335.

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer.

Leroux, B. and Rademacher, L. (2024). Euclidean distance compression via deep random features. *Advances in Neural Information Processing Systems*, 37:53261–53284.

Li, P., Church, K., and Hastie, T. (2008). One Sketch for All: Theory and Application of Conditional Random Sampling. *Advances in Neural Information Processing Systems*, 21.

Li, P., Hastie, T. J., and Church, K. W. (2006). Improving Random Projections Using Marginal Information. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, pages 635–649. Springer.

Li, X. and Li, P. (2022). C-MinHash: Improving minwise hashing with circulant permutation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12857–12887. PMLR.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

Meyer, R. A., Musco, C., Musco, C., and Woodruff, D. P. (2021). Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155. SIAM.

Nauman, F. and Herschel, M. (2022). *An Introduction to Duplicate Detection*. Springer Nature.

Oates, C. J., Girolami, M., and Chopin, N. (2017). Control Functionals for Monte Carlo Integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press.

Pratap, R. and Kulkarni, R. (2021). Variance Reduction in Frequency Estimators via Control Variates Method. In *Uncertainty in Artificial Intelligence*, pages 183–193. PMLR.

Pratap, R., Verma, B. D., and Kulkarni, R. (2021). Improving tug-of-war sketch using control-variates method. In *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA21)*, pages 66–76. SIAM.

Shao, J. (2003). *Mathematical Statistics*. Springer Science & Business Media.

South, L. F., Oates, C. J., Mira, A., and Drovandi, C. (2023). Regularized zero-variance control variates. *Bayesian Analysis*, 18(3):865–888.

Verma, B. D., Pratap, R., and Thakur, M. (2022). Variance Reduction in Feature Hashing using MLE and Control Variate Method. *Machine Learning*, 111(7):2631–2662.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009). Feature Hashing for Large Scale Multitask Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes. We give a clear description of exponential families, MLEs, and CVEs in Section 1, Section 2, Section 3, and Appendix B. We also know some readers may need a refresher on exponential families are, so we also recap in Appendix A.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes. We give an analysis of convergence in Appendix B, and experiments (including convergence properties) in Appendix C
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, see supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, we give the full set of assumptions for Theorem 4.
   (b) Complete proofs of all theoretical results. Yes. All proofs are in Appendix B.
   (c) Clear explanations of any assumptions. Yes, we give clear explanations in Section 2, Section 3, and Appendix B.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, the code is in the supplementary material. In fact, the instructions are on Page 24.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, they are described in Section 5 and Appendix C.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, they are described in Section 5 and Appendix C.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, they are described in Section 5 and Appendix C.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes, we cite authors of algorithms we used.
   (b) The license information of the assets, if applicable. Not applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not applicable.
   (d) Information about consent from data providers/curators. Not applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not applicable.

Supplementary Materials

A EXPONENTIAL FAMILY MODEL WITH BIVARIATE NORMAL

We give an example of the notation in Sections 1.1-2.2 using the zero-mean bivariate Normal. We use this distribution as several authors (Li et al., 2006; Kang, 2021; Pratap and Kulkarni, 2021; Verma et al., 2022) use either MLE or CVE in the context of feature hashing or similarity search, when **the underlying distribution is the zero-mean bivariate (or multivariate) Normal**.

The probability density function of the bivariate Normal with zero mean for n observations is

$$p(\mathcal{X}|\Sigma) = \frac{1}{(2\pi)^n|\Sigma|^{n/2}} \exp\left\{\sum_{i=1}^{n} -\frac{1}{2}\begin{pmatrix} x_{1_i} \\ x_{2_i} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} x_{1_i} \\ x_{2_i} \end{pmatrix}\right\}. \tag{6}$$

The parameters of the distribution are given by $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$, and we write $\vec{\nu} \equiv (\nu_1, \nu_2, \nu_3) = (\sigma_{11}, \sigma_{22}, \sigma_{12})$ noting that $\sigma_{12} = \sigma_{21}$. Equation (6) can be expressed in exponential family form as

$$\begin{aligned}
p(\mathcal{X}|\Sigma) &= \frac{\exp\left\{-\frac{\sum_{i=1}^{n}\left(x_{1_i}^2\sigma_{22}+x_{2_i}^2\sigma_{11}-2x_{1_i}x_{2_i}\sigma_{12}\right)}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\right\}}{(2\pi)^n|\Sigma|^{n/2}} \\
&= \frac{\exp\left\{-\frac{\sum_{i=1}^{n} x_{1_i}^2\sigma_{22}+x_{2_i}^2\sigma_{11}-2x_{1_i}x_{2_i}\sigma_{12}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\right\}}{\exp\left\{\frac{n}{2}\log(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\right\}} \frac{1}{(2\pi)^{\frac{n}{2}}} \\
&= \frac{\exp\left\{\begin{pmatrix} \frac{-\sigma_{22}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} \\ \frac{-\sigma_{11}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} \\ \frac{\sigma_{12}}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} \end{pmatrix}^T \begin{pmatrix} \sum_{i=1}^{n} x_{1_i}^2 \\ \sum_{i=1}^{n} x_{2_i}^2 \\ \sum_{i=1}^{n} x_{1_i}x_{2_i} \end{pmatrix}\right\}}{\exp\left\{\frac{n}{2}\left(\log(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\right)\right\}} \frac{1}{(2\pi)^{\frac{n}{2}}} \\
&= \frac{\exp\left\{n\begin{pmatrix} \frac{-\sigma_{22}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} \\ \frac{-\sigma_{11}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} \\ \frac{\sigma_{12}}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} \end{pmatrix}^T \begin{pmatrix} \frac{\sum_{i=1}^{n} x_{1_i}^2}{n} \\ \frac{\sum_{i=1}^{n} x_{2_i}^2}{n} \\ \frac{\sum_{i=1}^{n} x_{1_i}x_{2_i}}{n} \end{pmatrix}\right\}}{\exp\left\{n\left(\frac{1}{2}\log(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\right)\right\}} \frac{1}{(2\pi)^{\frac{n}{2}}} \\
&= \frac{\exp\left\{n\vec{\eta}^T\vec{y}\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g\left(\{\vec{x}_i\}_{i=1}^n\right)
\end{aligned}$$

where

$$\begin{aligned}
\vec{\eta}^T &= \left(\frac{-\sigma_{22}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}, \frac{-\sigma_{11}}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}, \frac{\sigma_{12}}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\right) \\
\vec{y}^T &= \left(\frac{\sum_{i=1}^{n} x_{1_i}^2}{n}, \frac{\sum_{i=1}^{n} x_{2_i}^2}{n}, \frac{\sum_{i=1}^{n} x_{1_i}x_{2_i}}{n}\right) \\
\vec{\nu}^T &= (\sigma_{11} \ , \ \sigma_{22} \ , \ \sigma_{12}) \\
&= \left(-\frac{2\eta_2}{4\eta_1\eta_2-\eta_3^2}, -\frac{2\eta_1}{4\eta_1\eta_2-\eta_3^2}, \frac{\eta_3}{4\eta_1\eta_2-\eta_3^2}\right) \\
\psi(\vec{\eta}) &= -\frac{1}{2}\log\left(4\eta_1\eta_2-\eta_3^2\right).
\end{aligned}$$

In practice, we do not need to express $\psi(\vec{\eta})$ in terms of $\vec{\eta}$ as $\vec{\mu}$ and $\mathbf{V}$ can be computed via the multivariate chain rule using partial derivatives since

$$\frac{\partial \psi(\vec{\eta})}{\partial \eta_i} = \sum_{j=1}^{p} \frac{\partial \psi(\vec{\eta})}{\partial \nu_j} \frac{\partial \nu_j}{\partial \eta_i} \; \forall \; 1 \leq i \leq p.$$

The mean vector $\vec{\mu}$ is computed via Theorem 2 with

$$\begin{aligned}
\frac{\partial \psi(\vec{\eta})}{\partial \eta_1} &= -\frac{2\eta_2}{4\eta_1\eta_2 - \eta_3^2} = \sigma_{11} \\
\frac{\partial \psi(\vec{\eta})}{\partial \eta_2} &= -\frac{2\eta_1}{4\eta_1\eta_2 - \eta_3^2} = \sigma_{22} \\
\frac{\partial \psi(\vec{\eta})}{\partial \eta_3} &= \frac{\eta_3}{4\eta_1\eta_2 - \eta_3^2} = \sigma_{12}
\end{aligned}$$

and they match the expectations of $\vec{y}$. V_n is analogously computed via Theorem 2 where

$$\mathbf{V}_n = \frac{1}{n} \begin{pmatrix} \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_1{}^2} & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_1 \partial \eta_2} & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_1 \partial \eta_3} \\ \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_2 \partial \eta_1} & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_2{}^2} & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_2 \partial \eta_3} \\ \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_3 \partial \eta_1} & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_3 \partial \eta_2} & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_3{}^2} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 2\sigma_{11}^2 & 2\sigma_{12}^2 & 2\sigma_{11}\sigma_{12} \\ 2\sigma_{12}^2 & 2\sigma_{22}^2 & 2\sigma_{22}\sigma_{12} \\ 2\sigma_{11}\sigma_{12} & 2\sigma_{22}\sigma_{12} & \sigma_{11}\sigma_{22} + \sigma_{12}^2 \end{pmatrix},$$

which matches the variance-covariances between $\vec{y}$.

The MLEs of $\vec{\eta}$ are given by $\vec{y}$ from the score function. To find the MLE of $\vec{\nu}$, we can use the invariance properties of MLE with the estimates $\hat{\eta}$. In the case of the bivariate Normal with zero mean, the MLE of $\vec{\eta}$ is exactly the same as the MLE of $\vec{\nu}$ (up to some permutation of its entries).

Suppose we know σ_{11}, σ_{22}, and want an MLE for σ_{12}. We compute the score function given by Equation (3) to get

$$\begin{aligned}
\frac{\partial l(\mathcal{X}|\vec{\eta}(\sigma_{12}))}{\partial \sigma_{12}} &= n \left(\frac{\partial \vec{\eta}(\sigma_{12})}{\partial \sigma_{12}} \right)^T (\vec{y} - \vec{\mu}) \\
&= n \begin{pmatrix} \frac{-\sigma_{22}\sigma_{12}}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2} \\ \frac{-\sigma_{11}\sigma_{12}}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2} \\ \frac{\sigma_{11}\sigma_{12} + \sigma_{12}^2}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2} \end{pmatrix}^T \left(\begin{pmatrix} \frac{\sum_{i=1}^n x_{1_i}^2}{n} \\ \frac{\sum_{i=1}^n x_{2_i}^2}{n} \\ \frac{\sum_{i=1}^n x_{1_i} x_{2_i}}{n} \end{pmatrix} - \begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{pmatrix} \right).
\end{aligned} \tag{7}$$

Setting the score function in Equation (7) to zero leads to

$$-\sigma_{22}\sigma_{12} \left(\frac{\sum_{i=1}^n x_{1_i}^2}{n} - \sigma_{11} \right) - \sigma_{11}\sigma_{12} \left(\frac{\sum_{i=1}^n x_{2_i}^2}{n} - \sigma_{22} \right) + (\sigma_{11}\sigma_{12} + \sigma_{12}^2) \left(\frac{\sum_{i=1}^n x_{1_i} x_{2_i}}{n} - \sigma_{12} \right) = 0 \tag{8}$$

where $\frac{n}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$ is factored out, since $\det(\Sigma) = \sigma_{11}\sigma_{22} - \sigma_{12}^2 > 0$, due to positive-definiteness of the covariance matrix.

The LHS of Equation 8 is a cubic in σ_{12}, and can be solved via root-finding methods to get $\hat{\sigma}_{12}$.

Using the differential relationship in Theorem 3, we have $\left(\frac{\partial \mu_1}{\partial \sigma_{12}}, \ldots, \frac{\partial \mu_3}{\partial \sigma_{12}} \right)^T = \frac{1}{n}(0, 0, 1)^T = \mathbf{V}_n \frac{\partial \vec{\eta}(\sigma_{12})}{\partial \sigma_{12}}$. From Equation (3), the Fisher Information with respect to σ_{12} is

$$\begin{aligned}
i_{\sigma_{12}} &= \mathbb{E} \left(n^2 \frac{\partial \vec{\eta}(\sigma_{12})}{\partial \sigma_{12}} \left(\vec{y} - \frac{\partial \psi(\vec{\eta})}{\partial \vec{\eta}} \right) \left(\vec{y} - \frac{\partial \psi(\vec{\eta})}{\partial \vec{\eta}} \right)^T \frac{\partial \vec{\eta}(\sigma_{12})}{\partial \sigma_{12}} \right) \\
&= n^2 \frac{\partial \vec{\eta}(\sigma_{12})}{\partial \sigma_{12}} \mathbf{V} \frac{\partial \vec{\eta}(\sigma_{12})}{\partial \sigma_{12}} = n \frac{\partial \eta_3}{\partial \sigma_{12}}.
\end{aligned}$$

This implies the distribution of the MLE for $\hat{\sigma}_{12}$ converges to $\mathcal{N}\left(\sigma_{12}, \frac{1}{n\frac{\partial \eta_3}{\partial \sigma_{12}}}\right)$, i.e. the asymptotic variance of $\hat{\sigma}_{12}$ is $\frac{(\sigma_{11}\sigma_{12}-\sigma_{12}^2)^2}{n(\sigma_{11}\sigma_{12}+\sigma_{12}^2)}$. Now suppose we know σ_{11}, σ_{22}, and want a CVE for σ_{12}. The CVE given by the sufficient statistics is

$$\begin{aligned}\hat{\sigma}_{12} &= \mathbb{E}\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{n} + c_1\left(\frac{\sum_{i=1}^n x_{1i}^2}{n} - \sigma_{11}\right) + c_2\left(\frac{\sum_{i=1}^n x_{2i}^2}{n} - \sigma_{22}\right)\right) \\ &= \mathbb{E}\left(y_3 + \sum_{i=1}^2 c_i(y_i - \mu_i)\right). \end{aligned} \tag{9}$$

The variance of $\hat{\sigma}_{12}$ in Equation (9) is given by

$$\text{Var}\left(y_3\right) + \sum_{i=1}^2 c_i^2 \text{Var}\left(y_i\right) + 2\sum_{i=1}^2 c_i \text{Cov}\left(y_i, y_3\right) + 2\sum_{i>j}^2 c_i c_j \text{Cov}\left(y_i, y_j\right),$$

and the optimal value of $\hat{c}_i$ given by $\widetilde{\mathbf{V}}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = -\vec{d}$ with $\widetilde{\mathbf{V}} = \begin{pmatrix} \text{Var}\left(y_1\right) & \text{Cov}\left(y_1, y_2\right) \\ \text{Cov}\left(y_2, y_1\right) & \text{Var}\left(y_2\right) \end{pmatrix}$ and $\vec{d} = \begin{pmatrix} \text{Cov}\left(y_1, y_3\right) \\ \text{Cov}\left(y_2, y_3\right) \end{pmatrix}$. Solving for $\hat{c}_1$ and $\hat{c}_2$ gives

$$\hat{c}_1 = -\frac{\sigma_{12}\sigma_{22}}{\sigma_{12}^2 + \sigma_{11}\sigma_{22}} \qquad \hat{c}_2 = -\frac{\sigma_{12}\sigma_{11}}{\sigma_{12}^2 + \sigma_{11}\sigma_{22}}.$$

and thus the CVE is given by

$$\hat{\sigma}_{12} = \mathbb{E}\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{n} - \frac{\sigma_{12}\sigma_{22}}{\sigma_{12}^2 + \sigma_{11}\sigma_{22}}\left(\frac{\sum_{i=1}^n x_{1i}^2}{n} - \sigma_{11}\right) - \frac{\sigma_{12}\sigma_{11}}{\sigma_{12}^2 + \sigma_{11}\sigma_{22}}\left(\frac{\sum_{i=1}^n x_{2i}^2}{n} - \sigma_{22}\right)\right) \tag{10}$$

The variance-covariance matrix $\mathbf{V}_n$ can be partitioned as

$$\mathbf{V}_n = \left(\begin{array}{c|c} \widetilde{\mathbf{V}} & \vec{d}^T \\ \hline \vec{d} & \text{Var}\left(y_3\right) \end{array}\right),$$

hence the variance given by the CVE is

$$\begin{aligned}\text{Var}\left(\hat{\sigma}_{12}\right) &= \text{Var}\left(y_3\right) - \vec{d}^T\widetilde{\mathbf{V}}^{-1}\vec{d} \\ &= \mathbf{V}_{3,3} - \mathbf{V}_{3,1:2}(\mathbf{V}_{1:2,1:2})^{-1}\mathbf{V}_{1:2,3} \\ &= \frac{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2}{n(\sigma_{11}\sigma_{22} + \sigma_{12}^2)} \end{aligned} \tag{11}$$

We make the following observations.

1. The asymptotic variance given by the MLE of $\hat{\sigma}_{12}$ is $\frac{(\sigma_{11}\sigma_{12}-\sigma_{12}^2)^2}{n(\sigma_{11}\sigma_{12}+\sigma_{12}^2)}$. However, Equation (11) also gives an identical variance.
2. The CVE for $\hat{\sigma}_{12}$ in Equation (10) contains σ_{12} in the optimal $\hat{c}_1$, $\hat{c}_2$, which we do not have, since the goal was to estimate σ_{12}. Substituting the empirical values of $\hat{\sigma}_{12}$ from the data would invalidate the variance calculations for the CVE for Equation (11).
3. This naturally leads to a question: *How can the asymptotic variance for the MLE be "correct" when n is large, but the same variance for the CVE be "wrong" (invalid) since $\hat{c}_1, \hat{c}_2$ uses the empirical $\hat{\sigma}_{12}$?*
4. We observe that if the expectation was removed from Equation (10), and we let

$$\sigma_{12} = \frac{\sum_{i=1}^n x_{1i}x_{2i}}{n} - \frac{\sigma_{12}\sigma_{22}}{\sigma_{12}^2 + \sigma_{11}\sigma_{22}}\left(\frac{\sum_{i=1}^n x_{1i}^2}{n} - \sigma_{11}\right) - \frac{\sigma_{12}\sigma_{11}}{\sigma_{12}^2 + \sigma_{11}\sigma_{22}}\left(\frac{\sum_{i=1}^n x_{2i}^2}{n} - \sigma_{22}\right), \tag{12}$$

collecting terms on one side of Equation 12 yields the same expression as Equation 8, up to scalar multiples.

5. Observation 1 seems to suggest that we can somehow make use of the CVE to get an estimate of $\hat{\sigma}_{12}$, even if $\hat{c}_1, \hat{c}_2$ contain σ_{12} since the MLE and CVE have the same (asymptotic) variance for $\hat{\sigma}_{12}$. Observation 4 somehow suggests a fixed point iteration (or similar) given the same terms in Equation 8 and Equation 12, which motivates our Algorithm 1.

In the main text of the paper, we will give conditions for the equivalence of variances for the estimators under MLE and CVE. We show that the structure of the CVE and MLE leads to a fixed point algorithm which converges to the required estimates. ionion

B PROOFS

B.1 PRELIMINARY LEMMA AND THEOREM

We give a lemma and a theorem which we use in the main text.

Lemma 1. *Suppose we have an invertible map between $\vec{\eta} \in \mathbb{R}^p$ and $\vec{\mu} \in \mathbb{R}^p$, and the Jacobian is given by* $\mathbf{V} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & \cdots & \frac{\partial \mu_1}{\partial \eta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_p}{\partial \eta_1} & \cdots & \frac{\partial \mu_p}{\partial \eta_p} \end{pmatrix}$. *Then* $\mathbf{V}^{-1} = \begin{pmatrix} \frac{\partial \eta_1}{\partial \mu_1} & \cdots & \frac{\partial \eta_1}{\partial \mu_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_p}{\partial \mu_1} & \cdots & \frac{\partial \eta_p}{\partial \mu_p} \end{pmatrix}$.

Proof. Consider $\mathbf{G} = \mathbf{V}\mathbf{V}^{-1}$. Then $\mathbf{G}_{ii} = \sum_{s=1}^{p} \frac{\partial \mu_i}{\partial \eta_s} \frac{\partial \eta_s}{\partial \mu_i} = \frac{\partial \mu_i}{\partial \mu_i} = 1$ and $\mathbf{G}_{ij} = \sum_{s=1}^{p} \frac{\partial \mu_i}{\partial \eta_s} \frac{\partial \eta_s}{\partial \mu_j} = \frac{\partial \mu_i}{\partial \mu_j} = 0$ for $i \neq j$ by the multivariate chain rule, implying $\mathbf{G} = \mathbf{I}$. □

Theorem 5. *(Schur Complement)*

Suppose we have an invertible block matrix $\mathbf{M} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$. *If* $\mathbf{A}$ *and* $\mathbf{D}$ *are invertible, then*

$$\mathbf{M}^{-1} = \left(\begin{array}{c|c} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ \hline -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{array}\right).$$

Proof. A proof can be found in Golub and Van Loan (2013). □

B.2 PROOF OF VARIANCE EQUIVALENCE BETWEEN MLE AND CVE

We now give a detailed proof of the variance equivalence.

Theorem 2. *Given n observations, the mean vector $\vec{\mu} \equiv \mathbb{E}(\vec{y})$ and the variance-covariance matrix $\mathbf{V}_n \equiv \mathrm{Cov}(\vec{y})$ for $\vec{y}$ in the exponential family is given by $\vec{\mu} = \left(\frac{\partial \psi(\vec{\eta})}{\partial \eta_1}, \ldots, \frac{\partial \psi(\vec{\eta})}{\partial \eta_p}\right)$ and $n\mathbf{V}_n = \mathbf{V}$ where* $\mathbf{V} = \begin{pmatrix} \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_1^2} & \cdots & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_1 \partial \eta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_p \eta_1} & \cdots & \frac{\partial^2 \psi(\vec{\eta})}{\partial \eta_p^2} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & \cdots & \frac{\partial \mu_1}{\partial \eta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_p}{\partial \eta_1} & \cdots & \frac{\partial \mu_p}{\partial \eta_p} \end{pmatrix}$.

Proof. Let $\mathcal{S}$ denote the support of $\mathcal{X}$, and write $y_i = y_i(\mathcal{X})$ below for ease of notation. Since

$$p(\mathcal{X} \mid \vec{\eta}) = \frac{\exp\left\{n\vec{\eta}^T \vec{y}(\mathcal{X})\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g(\mathcal{X}),$$

the normalizing condition gives

$$\int_{\mathcal{S}} \frac{\exp\left\{n\vec{\eta}^T \vec{y}(\mathcal{X})\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g(\mathcal{X}) \, \mathrm{d}\vec{x}_1 \cdots \mathrm{d}\vec{x}_n = 1.$$

Equivalently,

$$\int_{\mathcal{S}} \exp\left\{n\vec{\eta}^T \vec{y}(\mathcal{X})\right\} g(\mathcal{X}) \, \mathrm{d}\vec{x}_1 \cdots \mathrm{d}\vec{x}_n = \exp\left\{n\psi(\vec{\eta})\right\}. \tag{13}$$

Assume that $\vec{\eta}$ is in the interior of the natural parameter space. Then the derivative can be taken under the integral as the dominated convergence properties hold for exponential families. A proof of this can be found in Section 2.3 of Keener (2010). Differentiating Equation (13) with respect to η_i gives

$$\begin{aligned} \int_{\mathcal{S}} n y_i \exp\left\{n\vec{\eta}^T \vec{y}(\mathcal{X})\right\} g(\mathcal{X}) \, \mathrm{d}\vec{x}_1 \cdots \mathrm{d}\vec{x}_n &= n \exp\left\{n\psi(\vec{\eta})\right\} \frac{\partial \psi(\vec{\eta})}{\partial \eta_i} \\ \Rightarrow \quad \frac{\partial \psi(\vec{\eta})}{\partial \eta_i} &= \int_{\mathcal{S}} y_i \frac{\exp\left\{n\vec{\eta}^T \vec{y}(\mathcal{X})\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g(\mathcal{X}) \, \mathrm{d}\vec{x}_1 \cdots \mathrm{d}\vec{x}_n \\ &= \mathbb{E}(y_i). \end{aligned} \tag{14}$$

This gives the mean of y_i. Differentiating Equation (14) again with respect to η_i gives

$$\begin{aligned}\int_{\mathcal{S}} n^2 y_i^2 \exp\left\{n\vec{\eta}^T\vec{y}(\mathcal{X})\right\} g(\mathcal{X})\, \mathrm{d}\vec{x}_1 \cdots \mathrm{d}\vec{x}_n &= n\left(\exp\left\{n\psi(\vec{\eta})\right\} \frac{\partial^2\psi(\vec{\eta})}{\partial\eta_i^2} + n\exp\left\{n\psi(\vec{\eta})\right\}\left(\frac{\partial\psi(\vec{\eta})}{\partial\eta_i}\right)^2\right)\\ \Rightarrow \quad n\mathbb{E}\left(y_i^2\right) &= \frac{\partial^2\psi(\vec{\eta})}{\partial\eta_i^2} + n\left(\mathbb{E}\left(y_i\right)\right)^2\\ \Rightarrow \quad \frac{\partial^2\psi(\vec{\eta})}{\partial\eta_i^2} &= n\left(\mathbb{E}\left(y_i^2\right) - \mathbb{E}\left(y_i\right)^2\right) = n\mathrm{Var}\left(y_i\right).\end{aligned}$$

Similarly, differentiating Equation (14) with respect to η_j, where $j \neq i$, gives

$$\begin{aligned}\int_{\mathcal{S}} n^2 y_i y_j \exp\left\{n\vec{\eta}^T\vec{y}(\mathcal{X})\right\} g(\mathcal{X})\, \mathrm{d}\vec{x}_1 \cdots \mathrm{d}\vec{x}_n &= n\left(\exp\left\{n\psi(\vec{\eta})\right\} \frac{\partial^2\psi(\vec{\eta})}{\partial\eta_i\partial\eta_j} + n\exp\left\{n\psi(\vec{\eta})\right\}\frac{\partial\psi(\vec{\eta})}{\partial\eta_i}\frac{\partial\psi(\vec{\eta})}{\partial\eta_j}\right)\\ \Rightarrow \quad n\mathbb{E}\left(y_i y_j\right) &= \frac{\partial^2\psi(\vec{\eta})}{\partial\eta_i\partial\eta_j} + n\mathbb{E}\left(y_i\right)\mathbb{E}\left(y_j\right)\\ \Rightarrow \quad \frac{\partial^2\psi(\vec{\eta})}{\partial\eta_i\partial\eta_j} &= n\left(\mathbb{E}\left(y_i y_j\right) - \mathbb{E}\left(y_i\right)\mathbb{E}\left(y_j\right)\right) = n\mathrm{Cov}\left(y_i, y_j\right),\end{aligned}$$

which completes the proof. □

Theorem 4. *Let observations $\mathcal{X} := \{\vec{x}_i\}_{i=1}^n$ come from an exponential family which is regular and minimal, with ψ twice continuously differentiable and $\mathbf{V} = \nabla^2\psi(\vec{\eta}) \succ 0$ at the parameter value under consideration. Let $\nu_{\mathrm{E}} := \{\nu_i\}_{i=1}^t$ be the set of parameters to be estimated, and $\nu_{\mathrm{K}} := \{\nu_j\}_{j=t+1}^p$ be the set of parameters known. Suppose every $\mu_i \equiv \frac{\partial\psi(\vec{\eta})}{\partial\eta_i} = \mathbb{E}\left(y_i\right)$ is equal to ν_i. The asymptotic variance reduction given by the MLE of a linear combination of ν_{E} is equal to the variance reduction given by a CVE where components in $\vec{y}$ are used as control variates.*

Proof. Let $Y = \sum_{i=1}^t \alpha_i y_i$ for $\alpha_i \in \mathbb{R}$. Clearly, $\mathbb{E}\left(Y\right) = \mathbb{E}\left(\sum_{i=1}^t \alpha_i y_i\right) = \sum_{i=1}^t \alpha_i\nu_i$, and $\mathrm{Var}\left(Y\right) = \sum_{i=1}^t \alpha_i^2\mathrm{Var}\left(y_i\right) + 2\sum_{i,j}^t \alpha_i\alpha_j\mathrm{Cov}\left(y_i, y_j\right) = \vec{\alpha}^T(\mathbf{V}_n)_{1:t,1:t}\vec{\alpha}$, where $\vec{\alpha}^T = (\alpha_1, \ldots, \alpha_t)$ and $\mathbf{V}_n$ comes from Theorem 2.

Our proof strategy is to obtain expressions for the **variances** of the estimate of Y under the MLE and CVE, and show that they are equal.

The MLE for ν_{E} is asymptotically distributed as $\mathcal{N}\left(\nu_{\mathrm{E}}, i_{\nu_{\mathrm{E}}}^{-1}\right)$, so the asymptotic variance of Y is $\vec{\alpha}^T i_{\nu_{\mathrm{E}}}^{-1}\vec{\alpha}$.

We now compute $i_{\nu_{\mathrm{E}}}$ by substituting the score function from Equation (3):

$$\begin{aligned}i_{\nu_{\mathrm{E}}} &= \mathbb{E}\left(n^2\left(\frac{\partial\vec{\eta}(\nu_{\mathrm{E}})}{\partial\nu_{\mathrm{E}}}\right)^T\left(\vec{y} - \frac{\partial\psi(\vec{\eta})}{\partial\vec{\eta}}\right)\left(\vec{y} - \frac{\partial\psi(\vec{\eta})}{\partial\vec{\eta}}\right)^T\frac{\partial\vec{\eta}(\nu_{\mathrm{E}})}{\partial\nu_{\mathrm{E}}}\right)\\ &= n^2\left(\frac{\partial\vec{\eta}(\nu_{\mathrm{E}})}{\partial\nu_{\mathrm{E}}}\right)^T\mathbf{V}_n\frac{\partial\vec{\eta}(\nu_{\mathrm{E}})}{\partial\nu_{\mathrm{E}}}, \qquad (15)\end{aligned}$$

where Equation (15) follows from Theorem 2. Using the differential relationship in Theorem 3 and the fact that $\mu_i = \nu_i$, we have

$$\mathbf{V}_n\frac{\partial\vec{\eta}(\nu_{\mathrm{E}})}{\partial\nu_{\mathrm{E}}} = \frac{1}{n}\begin{pmatrix}\frac{\partial\mu_1}{\partial\nu_1} & \cdots & \frac{\partial\mu_1}{\partial\nu_t}\\ \vdots & \ddots & \vdots\\ \frac{\partial\mu_p}{\partial\nu_1} & \cdots & \frac{\partial\mu_p}{\partial\nu_t}\end{pmatrix} = \frac{1}{n}\begin{pmatrix}\mathbf{I}_{t\times t}\\ \mathbf{0}_{(p-t)\times t}\end{pmatrix},$$

where $\mathbf{I}_{t\times t}$ is the identity matrix. Therefore,

$$i_{\nu_{\mathrm{E}}} = n\left(\frac{\partial\vec{\eta}(\nu_{\mathrm{E}})}{\partial\nu_{\mathrm{E}}}\right)^T\begin{pmatrix}\mathbf{I}_t\\ \mathbf{0}_{(p-t)\times t}\end{pmatrix} = n\begin{pmatrix}\frac{\partial\eta_1}{\partial\mu_1} & \cdots & \frac{\partial\eta_1}{\partial\mu_t}\\ \vdots & \ddots & \vdots\\ \frac{\partial\eta_t}{\partial\mu_1} & \cdots & \frac{\partial\eta_t}{\partial\mu_t}\end{pmatrix}. \qquad (16)$$

By Equation (16), the asymptotic **variance** given by the MLE is

$$\mathbf{V}^{\mathrm{MLE}} = \vec{\alpha}^T i_{\nu_{\mathrm{E}}}^{-1} \vec{\alpha} = \vec{\alpha}^T \left[\frac{1}{n} \begin{pmatrix} \frac{\partial \eta_1}{\partial \mu_1} & \cdots & \frac{\partial \eta_1}{\partial \mu_t} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_t}{\partial \mu_1} & \cdots & \frac{\partial \eta_t}{\partial \mu_t} \end{pmatrix}^{-1} \right] \vec{\alpha}. \tag{17}$$

We now consider the CVE

$$Y + \sum_{j=t+1}^{p} \hat{c}_j (y_j - \nu_j),$$

where $\hat{c}_j$ are the optimal CV corrections. Partition $\mathbf{V}$, with entries $\mathbf{v}_{ij} = \frac{\partial \mu_i}{\partial \eta_j}$, as

$$\mathbf{V} = \left(\begin{array}{c|c} \mathbf{A}_{t\times t} & \mathbf{B}_{t\times(p-t)} \\ \hline \mathbf{B}^T_{(p-t)\times t} & \mathbf{D}_{(p-t)\times(p-t)} \end{array} \right). \tag{18}$$

Because $\mathbf{V}$ is positive definite by assumption, its principal submatrix $\mathbf{D} = \mathbf{V}_{(t+1):p,(t+1):p}$ is also positive definite and hence invertible.

The covariance matrix of the control variates $(y_{t+1}, \ldots, y_p)^T$ is $\frac{1}{n}\mathbf{D}$. Moreover,

$$\mathrm{Cov}\left((y_{t+1}, \ldots, y_p)^T, Y\right) = \frac{1}{n}\mathbf{B}^T\vec{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{t} \alpha_i \frac{\partial \mu_i}{\partial \eta_{t+1}}, \ldots, \sum_{i=1}^{t} \alpha_i \frac{\partial \mu_i}{\partial \eta_p} \right)^T.$$

Therefore, the optimal CV coefficients satisfy

$$\frac{1}{n}\mathbf{D}\vec{c} = -\frac{1}{n}\mathbf{B}^T\vec{\alpha} = -\frac{1}{n}\left(\sum_{i=1}^{t} \alpha_i \frac{\partial \mu_i}{\partial \eta_{t+1}}, \ldots, \sum_{i=1}^{t} \alpha_i \frac{\partial \mu_i}{\partial \eta_p} \right)^T, \tag{19}$$

or equivalently, $\vec{c} = -\mathbf{D}^{-1}\mathbf{B}^T\vec{\alpha}$. Substituting this expression of $\vec{c}$ in Equation (4), the **variance** of the CVE is

$$\mathbf{V}^{\mathrm{CVE}} = \vec{\alpha}^T(\mathbf{V}_n)_{1:t,1:t}\vec{\alpha} - \frac{1}{n}\vec{\alpha}^T\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T\vec{\alpha} = \vec{\alpha}^T\left(\frac{1}{n}\mathbf{A} - \frac{1}{n}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T\right)\vec{\alpha}. \tag{20}$$

To show that the asymptotic **variance** given by the MLE in Equation (17) equals the **variance** of the CVE in Equation (20), it remains to show that

$$\begin{pmatrix} \frac{\partial \eta_1}{\partial \mu_1} & \cdots & \frac{\partial \eta_1}{\partial \mu_t} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_t}{\partial \mu_1} & \cdots & \frac{\partial \eta_t}{\partial \mu_t} \end{pmatrix}^{-1} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T.$$

We partition $\mathbf{V}^{-1}$ as

$$\mathbf{V}^{-1} = \left(\begin{array}{c|c} \widetilde{\mathbf{A}}_{t\times t} & \widetilde{\mathbf{B}}_{t\times(p-t)} \\ \hline \widetilde{\mathbf{B}}^T_{(p-t)\times t} & \widetilde{\mathbf{D}}_{(p-t)\times(p-t)} \end{array} \right),$$

where $(\mathbf{V}^{-1})_{ij} = \frac{\partial \eta_i}{\partial \mu_j}$ by Lemma 1. By the Schur Complement (Theorem 5), the upper-left block of $\mathbf{V}^{-1}$ is $\widetilde{\mathbf{A}} = \left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T\right)^{-1}$.

Since $\mathbf{V}$ and $\mathbf{D}$ are positive definite, the Schur complement $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T$ is positive definite and therefore invertible. Hence,

$$\begin{pmatrix} \frac{\partial \eta_1}{\partial \mu_1} & \cdots & \frac{\partial \eta_1}{\partial \mu_t} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_t}{\partial \mu_1} & \cdots & \frac{\partial \eta_t}{\partial \mu_t} \end{pmatrix}^{-1} = \widetilde{\mathbf{A}}^{-1} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T.$$

Substituting this identity into Equation (17) gives $\mathbf{V}^{\mathrm{MLE}} = \mathbf{V}^{\mathrm{CVE}}$, completing the proof. □

Corollary 1. *Theorem 4 also holds when* $\vec{\mu} = \mathbf{A}\vec{\nu}$.

Proof. Let

$$p(\mathcal{X} \mid \vec{\nu}) \equiv \frac{\exp\left\{n\vec{\eta}^T\vec{y}\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g\left(\mathcal{X}\right) \tag{21}$$

be the probability density expressed in exponential family form. Without loss of generality, let $\nu_{\mathrm{E}} = \{\nu_i\}_{i=1}^{t}$, and $\nu_{\mathrm{K}} = \{\nu_j\}_{j=t+1}^{p}$, and let $\vec{\mu} = \mathbf{A}\vec{\nu}$ where $\mathbf{A}$ is invertible.

$\mathbf{A}$ being invertible implies ν_i can be written as a linear combination of μ_i, with coefficients given by $(\mathbf{A}^{-1}\vec{\mu})_i$. But since $\mu_i = \mathbb{E}\left(y_i\right)$, then

$$\nu_i = \sum_{j=1}^{p}(\mathbf{A}^{-1})_{ij}\mu_j = \sum_{j=1}^{p}(\mathbf{A}^{-1})_{ij}\mathbb{E}\left(y_j\right) = \mathbb{E}\left(\sum_{j=1}^{p}(\mathbf{A}^{-1})_{ij}y_j\right).$$

Given that the sufficient statistics $\{y_i\}_{i=1}^{p}$ are linearly independent, then $\{(\mathbf{A}^{-1}\vec{y})_i\}_{i=1}^{p}$ have to be linearly independent as well.

Now let $\widetilde{y} = \mathbf{A}^{-1}\vec{y}$, $\widetilde{\eta} = \mathbf{A}^T\vec{\eta}$ and $\widetilde{\psi}(\widetilde{\eta}) = \psi\left(\mathbf{A}^{-T}\widetilde{\eta}\right)$. Since $\vec{\mu} = \mathbf{A}\vec{\nu}$,

$$\mathbb{E}\left(\widetilde{\vec{y}}\right) = \mathbf{A}^{-1}\mathbb{E}\left(\vec{y}\right) = \mathbf{A}^{-1}\vec{\mu} = \vec{\nu}.$$

Furthermore, since $\mathbf{A}$ is invertible and $\{y_i\}_{i=1}^{p}$ are linearly independent, the components of $\widetilde{\vec{y}}$ are linearly independent. Therefore, Equation (21) can be re-expressed as

$$p(\mathcal{X} \mid \vec{\nu}) = \frac{\exp\left\{n\vec{\eta}^T\vec{y}\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g\left(\mathcal{X}\right) = \frac{\exp\left\{n\left(\mathbf{A}^T\vec{\eta}\right)^T\left(\mathbf{A}^{-1}\vec{y}\right)\right\}}{\exp\left\{n\psi(\vec{\eta})\right\}} g\left(\mathcal{X}\right) = \frac{\exp\left\{n\widetilde{\vec{\eta}}^T\widetilde{\vec{y}}\right\}}{\exp\left\{n\widetilde{\psi}\left(\widetilde{\eta}\right)\right\}} g\left(\mathcal{X}\right). \tag{22}$$

The control variates are $\widetilde{y}_j$, $(t+1) \le j \le p$, and therefore the corresponding CVE is

$$\mathbb{E}\left(\left(\sum_{i=1}^{t}\alpha_i\widetilde{y}_i\right) + \sum_{j=t+1}^{p} c_j\left(\widetilde{y}_j - \nu_j\right)\right).$$

Applying Theorem 4 to the exponential family in Equation (22) completes the proof. □

B.3 PROOF BEHIND ALGORITHM 1

Here, we show that the optimal CVE gives a fixed point iteration that converges to the MLE, and give the conditions for convergence.

In the proofs that follow, the notation $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ denote the blocks of $\mathbf{V}$ in Equation (18), evaluated at $\vec{\eta}(\nu_{\mathrm{E}})$.

We first give a lemma that gives the optimal CV coefficient.

Lemma 2. *Under the assumptions of Theorem 4, consider estimators of* ν_{E} *of the form* $\vec{y}_{1:t} + \mathbf{C}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)$, *where* $\mathbf{C} \in \mathbb{R}^{t\times(p-t)}$. *Let* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{D}$ *be given by the partition in Equation* (18). *The unique choice of* $\mathbf{C}$ *which minimizes* $\mathrm{Cov}\left(\vec{y}_{1:t} + \mathbf{C}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right), \vec{y}_{1:t} + \mathbf{C}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)\right)$ *in the positive semidefinite ordering is* $\mathbf{C} = -\mathbf{B}\mathbf{D}^{-1}$. *Hence, for* $1 \le i \le t$ *and* $t+1 \le j \le p$, *the optimal CV coefficient in the* i^{th} *fixed point equation is* $\hat{c}_{ij} = -\left(\mathbf{B}\mathbf{D}^{-1}\right)_{i,j-t}$.

Proof. Since ν_{K} is fixed and not random,

$$\mathrm{Cov}\left(\vec{y}_{1:t} + \mathbf{C}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right), \vec{y}_{1:t} + \mathbf{C}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)\right) = \mathrm{Cov}\left(\vec{y}_{1:t} + \mathbf{C}\vec{y}_{(t+1):p}, \vec{y}_{1:t} + \mathbf{C}\vec{y}_{(t+1):p}\right).$$

By the partition in Equation (18),

$$\mathrm{Cov}\left(\vec{y}_{1:t}, \vec{y}_{1:t}\right) = \frac{1}{n}\mathbf{A}, \quad \mathrm{Cov}\left(\vec{y}_{1:t}, \vec{y}_{(t+1):p}\right) = \frac{1}{n}\mathbf{B}, \quad \mathrm{Cov}\left(\vec{y}_{(t+1):p}, \vec{y}_{(t+1):p}\right) = \frac{1}{n}\mathbf{D}.$$

Therefore,

$$\mathrm{Cov}\left(\vec{y}_{1:t} + \mathbf{C}\vec{y}_{(t+1):p}, \vec{y}_{1:t} + \mathbf{C}\vec{y}_{(t+1):p}\right) = \frac{1}{n}\left(\mathbf{A} + \mathbf{C}\mathbf{B}^T + \mathbf{B}\mathbf{C}^T + \mathbf{C}\mathbf{D}\mathbf{C}^T\right).$$

Completing the matrix square gives

$$\mathbf{A} + \mathbf{C}\mathbf{B}^T + \mathbf{B}\mathbf{C}^T + \mathbf{C}\mathbf{D}\mathbf{C}^T = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T + \left(\mathbf{C} + \mathbf{B}\mathbf{D}^{-1}\right)\mathbf{D}\left(\mathbf{C} + \mathbf{B}\mathbf{D}^{-1}\right)^T.$$

Hence,

$$\begin{aligned}\mathrm{Cov}&\left(\vec{y}_{1:t} + \mathbf{C}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right), \vec{y}_{1:t} + \mathbf{C}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)\right) \\ &= \frac{1}{n}\left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T\right) + \frac{1}{n}\left(\mathbf{C} + \mathbf{B}\mathbf{D}^{-1}\right)\mathbf{D}\left(\mathbf{C} + \mathbf{B}\mathbf{D}^{-1}\right)^T. \end{aligned} \tag{23}$$

Since $\mathbf{D}$ is positive definite, the second term in Equation (23) is positive semidefinite and equals the zero matrix if and only if $\mathbf{C} + \mathbf{B}\mathbf{D}^{-1} = \mathbf{0}$. Therefore, the unique choice of $\mathbf{C}$ which minimizes the covariance matrix in the positive semidefinite ordering is $\mathbf{C} = -\mathbf{B}\mathbf{D}^{-1}$. For $1 \le i \le t$ and $t+1 \le j \le p$, the coefficient multiplying $y_j - \mu_j$ in the i^{th} CV expression is the $(i, j-t)^{\text{th}}$ entry of $\mathbf{C}$. Thus, $\hat{c}_{ij} = -\left(\mathbf{B}\mathbf{D}^{-1}\right)_{i,j-t}$ and we are done.

□

While Theorem 4 shows that the MLE and optimal CVE have the same variance reduction, Theorem 6 below shows that the optimal CV corrections themselves turn the likelihood score equations into fixed point equations, so solving the CV system is equivalent to finding a likelihood stationary point.

Theorem 6. *Under the assumptions of Theorem 4, let* $\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T$. *Then*

$$\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}} = n\mathbf{S}^{-1}\left[\vec{y}_{1:t} - \nu_{\mathrm{E}} - \mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)\right].$$

Consequently, $\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}} = \vec{0}$ *if and only if* $\nu_{\mathrm{E}} = \vec{y}_{1:t} - \mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)$. *Equivalently, for every* $1 \le i \le t$, $\nu_i = y_i + \sum_{j=t+1}^{p} \hat{c}_{ij}(\nu_1, \ldots, \nu_t)(y_j - \mu_j)$ *where* $\hat{c}_{ij}(\nu_1, \ldots, \nu_t) = -\left(\mathbf{B}\mathbf{D}^{-1}\right)_{i,j-t}$.

Proof. Since $\mu_i = \nu_i$ and ν_{K} is fixed, $\frac{\partial \vec{\mu}}{\partial \nu_{\mathrm{E}}} = \begin{pmatrix} \mathbf{I}_{t\times t} \\ \hline \mathbf{0}_{(p-t)\times t} \end{pmatrix}$. By Theorem 3, $\mathbf{V}\frac{\partial \vec{\eta}(\nu_{\mathrm{E}})}{\partial \nu_{\mathrm{E}}} = \frac{\partial \vec{\mu}}{\partial \nu_{\mathrm{E}}}$. Therefore, $\frac{\partial \vec{\eta}(\nu_{\mathrm{E}})}{\partial \nu_{\mathrm{E}}} = \mathbf{V}^{-1}\begin{pmatrix} \mathbf{I}_{t\times t} \\ \hline \mathbf{0}_{(p-t)\times t} \end{pmatrix}$. Let $\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T$. By Theorem 5,

$$\mathbf{V}^{-1} = \left(\begin{array}{c|c} \mathbf{S}^{-1} & -\mathbf{S}^{-1}\mathbf{B}\mathbf{D}^{-1} \\ \hline -\mathbf{D}^{-1}\mathbf{B}^T\mathbf{S}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{B}^T\mathbf{S}^{-1}\mathbf{B}\mathbf{D}^{-1} \end{array}\right).$$

Thus,

$$\frac{\partial \vec{\eta}(\nu_{\mathrm{E}})}{\partial \nu_{\mathrm{E}}} = \begin{pmatrix} \mathbf{S}^{-1} \\ \hline -\mathbf{D}^{-1}\mathbf{B}^T\mathbf{S}^{-1} \end{pmatrix}.$$

Using Equation (3),

$$\begin{aligned}\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}} &= n\left(\frac{\partial \vec{\eta}(\nu_{\mathrm{E}})}{\partial \nu_{\mathrm{E}}}\right)^T(\vec{y} - \vec{\mu}) \\ &= n\begin{pmatrix} \mathbf{S}^{-1} \\ \hline -\mathbf{D}^{-1}\mathbf{B}^T\mathbf{S}^{-1} \end{pmatrix}^T\begin{pmatrix} \vec{y}_{1:t} - \nu_{\mathrm{E}} \\ \hline \vec{y}_{(t+1):p} - \nu_{\mathrm{K}} \end{pmatrix} \\ &= n\mathbf{S}^{-1}\left(\vec{y}_{1:t} - \nu_{\mathrm{E}}\right) - n\mathbf{S}^{-1}\mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right) \\ &= n\mathbf{S}^{-1}\left[\vec{y}_{1:t} - \nu_{\mathrm{E}} - \mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)\right]. \end{aligned} \tag{24}$$

Since $\mathbf{S}$ is invertible, Equation (24) implies that $\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}} = \vec{0}$ if and only if $\vec{y}_{1:t} - \nu_{\mathrm{E}} - \mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right) = \mathbf{0}$. Making ν_{E} the subject gives $\nu_{\mathrm{E}} = \vec{y}_{1:t} - \mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)$. By Lemma 2, $\hat{c}_{ij} = -\left(\mathbf{B}\mathbf{D}^{-1}\right)_{i,j-t}$. Finally, taking the i^{th} coordinate of the preceding vector equation gives

$$\nu_i = y_i + \sum_{j=t+1}^{p} \hat{c}_{ij}(\nu_1, \ldots, \nu_t)(y_j - \mu_j).$$

□

Corollary 2 now states that solving the fixed point equations yields a likelihood stationary point. If the likelihood has a unique stationary point that is the MLE, this stationary point is the MLE.

Corollary 2. *Suppose that* $\nu_{\mathrm{E}}^{\star}$ *satisfies*

$$\nu_i^{\star} = y_i + \sum_{j=t+1}^{p} \hat{c}_{ij}(\nu_1^{\star}, \ldots, \nu_t^{\star})(y_j - \mu_j), \quad 1 \leq i \leq t. \tag{25}$$

Then

$$\left.\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}}\right|_{\nu_{\mathrm{E}} = \nu_{\mathrm{E}}^{\star}} = \vec{0}.$$

Conversely, every interior solution of

$$\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}} = \vec{0}$$

satisfies Equation (25)*. Therefore, the fixed points of the CV expressions are exactly the interior likelihood stationary points. If Algorithm 1 converges to a fixed point and the likelihood has a unique stationary point that is the MLE, then its limit is the MLE of* ν_{E}*.*

Proof. Suppose that $\nu_{\mathrm{E}}^{\star}$ satisfies Equation (25). By the definition of $\hat{c}_{ij}$ in Theorem 6, stacking the t coordinate equations gives

$$\nu_{\mathrm{E}}^{\star} = \vec{y}_{1:t} - \mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right),$$

where $\mathbf{B}$ and $\mathbf{D}$ are evaluated at $\vec{\eta}(\nu_{\mathrm{E}}^{\star})$. By Theorem 6,

$$\left.\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}}\right|_{\nu_{\mathrm{E}} = \nu_{\mathrm{E}}^{\star}} = \vec{0}.$$

Thus, every fixed point of the CV expressions is a likelihood stationary point. Conversely, suppose that $\nu_{\mathrm{E}}^{\star}$ is an interior solution of

$$\left.\frac{\partial l(X \mid \vec{\eta}(\nu_{\mathrm{E}}))}{\partial \nu_{\mathrm{E}}}\right|_{\nu_{\mathrm{E}} = \nu_{\mathrm{E}}^{\star}} = \vec{0}.$$

By Theorem 6, $\nu_{\mathrm{E}}^{\star} = \vec{y}_{1:t} - \mathbf{B}\mathbf{D}^{-1}\left(\vec{y}_{(t+1):p} - \nu_{\mathrm{K}}\right)$. Taking the i^{th} coordinate and using $\hat{c}_{ij} = -\left(\mathbf{B}\mathbf{D}^{-1}\right)_{i,j-t}$ gives

$$\nu_i^{\star} = y_i + \sum_{j=t+1}^{p} \hat{c}_{ij}(\nu_1^{\star}, \ldots, \nu_t^{\star})(y_j - \mu_j), \quad 1 \leq i \leq t.$$

Therefore, every interior likelihood stationary point is a fixed point of the CV expressions.

If Algorithm 1 converges to a fixed point and the likelihood has a unique stationary point that is the MLE, then its limit is the MLE of ν_{E}. □

Finally, Theorem 7 gives a condition for Algorithm 1 to converge.

Theorem 7. *Suppose that $\nu_{\mathrm{E}}^\star$ satisfies Equation (25). For $1 \le i \le t$, define*

$$g_i(\nu_1, \ldots, \nu_t) = y_i + \sum_{j=t+1}^{p} \hat{c}_{ij}(\nu_1, \ldots, \nu_t)(y_j - \mu_j). \tag{26}$$

Suppose that $g_1, \ldots, g_t$ are continuously differentiable in a neighborhood of $\nu_{\mathrm{E}}^\star$. Let $\mathbf{G}$ denote one complete sequential update of Algorithm 1. That is, for $\nu_{\mathrm{E}} = (\nu_1, \ldots, \nu_t)^T$,

$$\mathbf{G}(\nu_{\mathrm{E}}) = \begin{pmatrix} g_1(\nu_1, \ldots, \nu_t) \\ g_2(g_1(\nu_1, \ldots, \nu_t), \nu_2, \ldots, \nu_t) \\ \vdots \\ g_t(\nu_1^+, \ldots, \nu_{t-1}^+, \nu_t) \end{pmatrix}, \tag{27}$$

where $\nu_1^+, \ldots, \nu_{t-1}^+$ denote the updated values obtained earlier in the same sequential iteration. Let

$$\mathbf{J}_{\mathbf{G}}(\nu_{\mathrm{E}}^\star) = \left(\frac{\partial G_i}{\partial \nu_j}(\nu_{\mathrm{E}}^\star) \right)_{1 \le i,j \le t} \tag{28}$$

denote the Jacobian matrix of $\mathbf{G}$ evaluated at $\nu_{\mathrm{E}}^\star$. If

$$\rho\left(\mathbf{J}_{\mathbf{G}}(\nu_{\mathrm{E}}^\star)\right) < 1 \tag{29}$$

where $\rho(\cdot)$ denotes the spectral radius, then there exists a neighborhood $\mathcal{N}$ of $\nu_{\mathrm{E}}^\star$ such that every initial estimate $\widehat{\nu_{\mathrm{E}}}^{(1)} \in \mathcal{N}$ produces iterates satisfying

$$\widehat{\nu_{\mathrm{E}}}^{(r)} \to \nu_{\mathrm{E}}^\star \quad \text{as } r \to \infty. \tag{30}$$

Proof. Since $\nu_{\mathrm{E}}^\star$ satisfies Equation (25), $\nu_i^\star = g_i(\nu_1^\star, \ldots, \nu_t^\star), 1 \le i \le t$. Therefore, one complete sequential update beginning at $\nu_{\mathrm{E}}^\star$ leaves every coordinate unchanged. Hence, $\mathbf{G}(\nu_{\mathrm{E}}^\star) = \nu_{\mathrm{E}}^\star$.

Now suppose that $\rho\left(\mathbf{J}_{\mathbf{G}}(\nu_{\mathrm{E}}^\star)\right) < 1$. Choose q such that $\rho\left(\mathbf{J}_{\mathbf{G}}(\nu_{\mathrm{E}}^\star)\right) < q < 1$. Then there exists a vector norm $\|\cdot\|_\star$ such that $\left\|\mathbf{J}_{\mathbf{G}}(\nu_{\mathrm{E}}^\star)\right\|_\star < q < 1$.

Since $g_1, \ldots, g_t$ are continuously differentiable in a neighborhood of $\nu_{\mathrm{E}}^\star$, $\mathbf{G}$ is continuously differentiable in a neighborhood of $\nu_{\mathrm{E}}^\star$. Therefore, there exists $r > 0$ such that $\mathcal{N} = \left\{\nu_{\mathrm{E}} : \left\|\nu_{\mathrm{E}} - \nu_{\mathrm{E}}^\star\right\|_\star \le r\right\}$ is contained in that neighborhood and $\sup_{\nu_{\mathrm{E}} \in \mathcal{N}} \left\|\mathbf{J}_{\mathbf{G}}(\nu_{\mathrm{E}})\right\|_\star \le q < 1$.

For any $\nu_{\mathrm{E}}, \widetilde{\nu_{\mathrm{E}}} \in \mathcal{N}$, the multivariate mean value inequality Ortega and Rheinboldt (1970) gives

$$\left\|\mathbf{G}(\nu_{\mathrm{E}}) - \mathbf{G}(\widetilde{\nu_{\mathrm{E}}})\right\|_\star \le q \left\|\nu_{\mathrm{E}} - \widetilde{\nu_{\mathrm{E}}}\right\|_\star.$$

Hence, for every $\nu_{\mathrm{E}} \in \mathcal{N}$,

$$\left\|\mathbf{G}(\nu_{\mathrm{E}}) - \nu_{\mathrm{E}}^\star\right\|_\star = \left\|\mathbf{G}(\nu_{\mathrm{E}}) - \mathbf{G}(\nu_{\mathrm{E}}^\star)\right\|_\star \le q \left\|\nu_{\mathrm{E}} \nu_{\mathrm{E}}^\star\right\|_\star \le qr < r.$$

Therefore, $\mathbf{G}(\mathcal{N}) \subseteq \mathcal{N}$, and $\mathbf{G}$ is a contraction from $\mathcal{N}$ to itself. Since $\mathbf{G}(\nu_{\mathrm{E}}^\star) = \nu_{\mathrm{E}}^\star$, the contraction mapping theorem Ortega and Rheinboldt (1970) implies that every initial estimate $\widehat{\nu_{\mathrm{E}}}^{(1)} \in \mathcal{N}$ produces iterates satisfying $\widehat{\nu_{\mathrm{E}}}^{(r)} \to \nu_{\mathrm{E}}^\star$ as $r \to \infty$.

□

B.4 DISCUSSION OF CONVERGENCE AND VARIANCE OF SMALL K

We give plots (Figure 14 and Figure 15) estimating the rate of convergence in Appendix C for feature hashing and random projections, and getting a bound on the rate of convergence would be a future open question for different types of exponential families.

In the case of the bivariate Normal (feature hashing and random projections), the form of the MLE is a cubic, and hence there is a possibility of the MLE (or Algorithm 1) converging to the wrong root. However, Lemma 4 in (Li et al., 2006) gives an upper bound the probability of the cubic having multiple real roots (which decreases exponentially fast as k increases). Our experiments in Appendix C (in particular Table 1 and Table 3) show that even with $k = 10$, the proportion of cubics with three real roots is negligible, and with $k = 20$, there are no cubics with three real roots.

On the other hand, if applying Algorithm 1 gives a closed form solution such as an unbiased estimator for the trace as discussed in Section 4 and Appendix D for Hutchinson's Trace Estimator, then we can just do standard variance analysis on the derived estimator as shown in Appendix D.

C SUPPLEMENT TO EXPERIMENTS

C.1 LAYOUT OF OUR SUPPLEMENTARY MATERIAL AND CODE FILES

We first describe the structure of code in our folder of supplementary material. The folder: `plots in paper` gives the plots that we use in paper in eps and pdf format, for ease of comparison.

The other two folders, `feature hashing` and `random projection` contain the code to generate all the plots in our paper.

There are three Matlab code files which we describe in detail below. The three files

- `newton_raphson_cubic_vectorized.m`
- `secant_cubic_vectorized.m`
- `cv_vectorized.m`

compute the estimates of $\langle \vec{x}_1, \vec{x}_2 \rangle$ under the respective algorithms, with 10 iterations/update steps till convergence at $\epsilon < 0.0001$. We pick 10 iterations/update steps till convergence since most estimates took at most 7 iterations/update steps till convergence, with remaining estimates failing to converge (even with 100 iterations/update steps) ; i.e. $\|f_{n+1} - f_n\|$ is not (almost always) monotonically decreasing.

We use the same vectorized procedure across all three Matlab files ; where only the update step changes. Algorithm 1 via the CVE is given by

$$f_{n+1} = \frac{\sum_{s=1}^{k} v_{1s} v_{2s}}{k} - \frac{f_n \left(\|\vec{x}_2\|^2 \left(\frac{\sum_{s=1}^{k} v_{1s}^2}{k} - \|\vec{x}_1\|^2 \right) + \|\vec{x}_1\|^2 \left(\frac{\sum_{s=1}^{k} v_{2s}^2}{k} - \|\vec{x}_2\|^2 \right) \right)}{f_n^2 + \|\vec{x}_1\|^2 \|\vec{x}_2\|^2}.$$

All three files do the following two checks after the 10 iterations/update steps to maintain consistency and to mimic what would have been done in an actual (non-experimental) situation.

1. If there was no convergence after 10 iterations/update steps, we set the estimate to be the initial estimate given by feature hashing / random projection.
2. If the converged estimate was greater than $\sqrt{\|\vec{x}_1\|^2 \|\vec{x}_2\|^2}$, we set it to be $\sqrt{\|\vec{x}_1\|^2 \|\vec{x}_2\|^2}$. Equivalently, if the converged estimate was lesser than $-\sqrt{\|\vec{x}_1\|^2 \|\vec{x}_2\|^2}$, we set it to be $-\sqrt{\|\vec{x}_1\|^2 \|\vec{x}_2\|^2}$. The reason for this comes from the cosine rule since

$$\cos(\theta) = \frac{\langle \vec{x}_1, \vec{x}_2 \rangle}{\sqrt{\|\vec{x}_1\|^2 \|\vec{x}_2\|^2}}$$

and as we would not know θ *a priori* (since we are estimating the inner product),

$$\begin{array}{rccccc} & -1 & \leq & \cos(\theta) & \leq & 1 \\ \Rightarrow & -\sqrt{\|\vec{x}_1\|^2 \|\vec{x}_2\|^2} & \leq & \langle \vec{x}_1, \vec{x}_2 \rangle & \leq & \sqrt{\|\vec{x}_1\|^2 \|\vec{x}_2\|^2}. \end{array}$$

The motivation for providing these three files is to show the difference (in implementation) for these three algorithms, since our experiments focus on comparing the performance of two common root-finding algorithms to recover the estimates of $\hat{\eta}$ under MLE, and `CV-FP` given by Algorithm 1 in our paper.

We stress that our paper demonstrates the performance of Algorithm 1. Therefore, we are not aiming to get improved performance on benchmarked datasets, or outperform state-of-the-art algorithms.

C.2 FEATURE HASHING: FASTER CONVERGENCE AND BETTER STABILITY FOR THE MLE

The feature hashing algorithm (Weinberger et al., 2009) is a widely used method to quickly estimate inner products between high dimensional vectors $\vec{x}_i, \vec{x}_j$. It is a hashing-based dimension reduction algorithm that

compresses high-dimensional data points into low-dimensional points so that the estimated pairwise inner product of the low-dimensional data points gives a close approximation of the corresponding pairwise inner product of the original high-dimensional data points.

Let $\mathbf{X}_{n\times p}$ be a data matrix, $h : [p] \mapsto [k]$ and $\phi : [p] \mapsto \{+1, -1\}$ be hash functions from a 2-wise universal hash family. Suppose $\vec{x}_i$ is the i-th row of the data matrix X and $\vec{v}_i$ be its k-dimensional sketch obtained using the feature hashing method. Then for any $s \in [k]$,

$$v_{is} = \sum_{t=1 \ s.t. \ h(t)=s}^{p} \phi(t)\, x_{it}. \tag{31}$$

Let $\mathbf{R}_{n\times k}$ be a matrix defined as $\mathbf{R} = \mathbf{DP}$, where $\mathbf{D}_{n\times n}$ is a diagonal matrix with i.i.d. ± 1 random entries and $\mathbf{P}_{n\times k}$ is a random matrix with exactly one non-zero entry at a random position per row, i.e. $\mathbf{P}_{i,:} \sim \text{Uniform}\{\vec{e}_1, \ldots, \vec{e}_p\}$, where for $t \in [p]$, $\vec{e}_t$ represents a standard basis of the p-dimensional space. Then, the feature hashing operation on the data matrix $\mathbf{X}$ can also be visualized as $\mathbf{V}_{n\times k} = \mathbf{XR}$. Suppose $\vec{v}_i, \vec{v}_j \in \mathbb{R}^k$ are the sketch vectors of the i-th and j-th row of the data matrix $\mathbf{X}$ (i.e. $\vec{x}_i$ and $\vec{x}_j$) obtained using feature hashing, respectively. Then $\langle \vec{v}_i, \vec{v}_j \rangle$ gives an unbiased estimate of $\langle \vec{x}_i, \vec{x}_j \rangle$.

Verma et al. (2022) used Lyapunov's Central Limit Theorem to show that for any two rows $\vec{v}_i, \vec{v}_j$ in the matrix $\mathbf{V}$, $(\vec{v}_i, \vec{v}_j) = (v_{i1,}, \ldots, v_{ik}, v_{j1}, \ldots, v_{jk})$ asymptotically follows the multivariate Normal distribution with

$$\begin{pmatrix} \vec{v}_i \\ \vec{v}_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \frac{1}{k} \begin{pmatrix} \|\vec{x}_i\|^2 & \cdots & 0 & \langle \vec{x}_i, \vec{x}_j \rangle & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \|\vec{x}_i\|^2 & 0 & \cdots & \langle \vec{x}_i, \vec{x}_j \rangle \\ \langle \vec{x}_i, \vec{x}_j \rangle & \cdots & 0 & \|\vec{x}_j\|^2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \langle \vec{x}_i, \vec{x}_j \rangle & 0 & \cdots & \|\vec{x}_j\|^2 \end{pmatrix} \right).$$

Verma et al. (2022) then used the asymptotic normality of $(\vec{v}_i, \vec{v}_j)$ and the fact that if the original ℓ_2 norm of the vectors $\vec{x}_1, \ldots, \vec{x}_n$ are known (as also considered in Li et al. (2006) in case of random projection) to construct an MLE to estimate $\langle \vec{x}_i, \vec{x}_j \rangle$. Their MLE estimate of the $\langle \vec{x}_i, \vec{x}_j \rangle$ corresponds to finding the root of the following cubic equation:

$$\lambda^3 - \left(\sum_{s=1}^{k} v_{is} v_{js} \right) \lambda^2 + \left(\|\vec{x}_i\|^2 \left(\sum_{s=1}^{k} v_{js}^2 \right) + \|\vec{x}_j\|^2 \left(\sum_{s=1}^{k} v_{is}^2 \right) - \|\vec{x}_i\|^2 \|\vec{x}_j\|^2 \right) \lambda$$
$$- \|\vec{x}_i\|^2 \|\vec{x}_j\|^2 \left(\sum_{s=1}^{k} v_{is} v_{js} \right) = 0. \tag{32}$$

However, instead of showing that $(\vec{v}_i, \vec{v}_j)$ follows a multivariate Normal distribution, applying Lyapunov's Central Limit Theorem shows that the distribution of the tuple (v_{is}, v_{js}), $1 \le s \le k$ is bivariate Normal with

$$\begin{pmatrix} v_{is} \\ v_{js} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{k} \begin{pmatrix} \|\vec{x}_i\|^2 & \langle \vec{x}_i, \vec{x}_j \rangle \\ \langle \vec{x}_i, \vec{x}_j \rangle & \|\vec{x}_j\|^2 \end{pmatrix} \right),$$

and hence an MLE is constructed to estimate $\langle \vec{v}_i, \vec{v}_j \rangle$. The probability density function of the distribution of $\{(v_{is}, v_{js})\}_{i=1}^{k}$ is given by

$$\frac{\exp\left\{ k \begin{pmatrix} \frac{-\|\vec{x}_j\|^2}{2(\|\vec{x}_i\|^2\|\vec{x}_j\|^2 - \langle \vec{x}_i, \vec{x}_j \rangle^2)} \\ \frac{-\|\vec{x}_i\|^2}{2(\|\vec{x}_i\|^2\|\vec{x}_j\|^2 - \langle \vec{x}_i, \vec{x}_j \rangle^2)} \\ \frac{\langle \vec{x}_i, \vec{x}_j \rangle}{(\|\vec{x}_i\|^2 \sigma_{22} - \langle \vec{x}_i, \vec{x}_j \rangle^2)} \end{pmatrix}^T \begin{pmatrix} \sum_{s=1}^{k} v_{is}^2 \\ \sum_{s=1}^{k} v_{js}^2 \\ \sum_{s=1}^{k} v_{is} v_{js} \end{pmatrix} \right\}}{\exp\left\{ k \left(\frac{1}{2} \log(\|\vec{x}_i\|^2 \|\vec{x}_j\|^2 - \langle \vec{x}_i, \vec{x}_j \rangle^2) \right) \right\}} \frac{1}{(2\pi)^{\frac{k}{2}}}$$

and the conditions in Theorem 4 hold. Hence, the asymptotic variance given by the MLE of $\langle \vec{x}_i, \vec{x}_j \rangle$ must be equivalent to the variance of the CVE.

The control variate estimator to estimate $\langle \vec{x}_i, \vec{x}_j \rangle$ is given by

$$\langle \vec{x}_i, \vec{x}_j \rangle = \mathbb{E}\left(\sum_{s=1}^{k} v_{is} v_{js} + \hat{c}_1 \left(\sum_{s=1}^{k} v_{is}^2 - \|\vec{x}_i\|^2 \right) + \hat{c}_2 \left(\sum_{s=1}^{k} v_{js}^2 - \|\vec{x}_j\|^2 \right) \right), \tag{33}$$

with

$$\hat{c}_1 = -\frac{\langle \vec{x}_i, \vec{x}_j \rangle \|\vec{x}_j\|^2}{\langle \vec{x}_i, \vec{x}_j \rangle^2 + \|\vec{x}_i\|^2 \|\vec{x}_j\|^2} \qquad \hat{c}_2 = -\frac{\langle \vec{x}_i, \vec{x}_j \rangle \|\vec{x}_i\|^2}{\langle \vec{x}_i, \vec{x}_j \rangle^2 + \|\vec{x}_i\|^2 \|\vec{x}_j\|^2}.$$

Since $\langle \vec{x}_i, \vec{x}_j \rangle$ is what we want to estimate, yet appears in $\hat{c}_1, \hat{c}_2$, Algorithm 1 gives an algorithm of the form

$$\begin{aligned} f_{n+1} &= \sum_{s=1}^{k} v_{is} v_{js} - \frac{f_n \|\vec{x}_j\|^2}{f_n^2 + \|\vec{x}_i\|^2 \|\vec{x}_j\|^2} \left(\sum_{s=1}^{k} v_{is}^2 - \|\vec{x}_i\|^2 \right) - \frac{f_n \|\vec{x}_i\|^2}{f_n^2 + \|\vec{x}_i\|^2 \|\vec{x}_j\|^2} \left(\sum_{s=1}^{k} v_{js}^2 - \|\vec{x}_j\|^2 \right) \\ &= \sum_{s=1}^{k} v_{is} v_{js} - \frac{f_n \left(\|\vec{x}_j\|^2 \left(\sum_{s=1}^{k} v_{is}^2 - \|\vec{x}_i\|^2 \right) + \|\vec{x}_i\|^2 \left(\sum_{s=1}^{k} v_{js}^2 - \|\vec{x}_j\|^2 \right) \right)}{f_n^2 + \|\vec{x}_i\|^2 \|\vec{x}_j\|^2} \end{aligned} \tag{34}$$

where we set $f_1 = \sum_{s=1}^{k} v_{is} v_{js}$.

Note that the control variate estimate given in Equation (33) is a generic version of the control variate estimate proposed by Verma et al. (2022) and converges to their estimate when $\hat{c}_1 = \hat{c}_2$.

We use the following experimental setup to demonstrate the practical applications of the equivalence between CVE and MLE using the feature hashing algorithm.

Hardware Description: We performed the experiments on a machine having the following configuration: CPU: Intel(R) Core(TM) i7-8750H CPU @ 2.21GHz x 6; Memory: 16 GB; OS: Windows 10.

Dataset: We randomly generate vector pairs $\vec{x}_1, \vec{x}_2 \in \mathbb{R}^{100,000}$ with varying squared norms and angles between them. Our experimental study involves the ratios $r \in \{0.1, 0.5, 1, 2, 10\}$ where $\|\vec{x}_1\|^2 = r\|\vec{x}_2\|^2$, and angles $\theta \in \{\frac{\pi}{12}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \frac{11\pi}{12}\}$.

We generate feature hashing sketches, denoted as $\vec{v}_i$ and $\vec{v}_j$, for the vector pairs $\vec{x}_i$ and $\vec{x}_j$, respectively, for different sketch sizes (k) ranging from 1 to 100 using Equation (31). We then compute the estimate of the inner product as $\langle \vec{x}_i, \vec{x}_j \rangle \approx \sum_{s=1}^{k} v_{is} v_{js}$ and call it the baseline estimate. Further, we compute estimates of the inner product using: a) MLE (Verma et al., 2022) (Equation (32)) via Newton Raphson (`MLE-NR`), b) MLE (Equation (32)) via the Secant method (`MLE-Secant`), c) CV (Equation (33) where the initial estimate $\sum_{s=1}^{k} v_{is} v_{js}$is used as part of $\hat{c}_i$ (`CV-Init`), d) CV (Equation (33)) where the empirical covariances and variances are found for $\hat{c}_i$ (`CV-Emp`), and e) our CVE (Algorithm 1) (`CV-FP`). The MLE formulation implies that all five methods are finding the root of a cubic. We repeat this procedure using different hash functions (or projection matrix $\mathbf{R}$) $10,000$ times and compute the corresponding inner product estimates. We also record the number of updates each method took to converge to the root and the number of cubic equations with three real roots over $10,000$ iterations.

We use the following metrics to evaluate the performance of the aforementioned five methods:

(i) **Mean Square Error (MSE) w.r.t. k over $10,000$ iterations across various combinations of r and θ:** A lower MSE indicates better performance of the method. We compute the MSE of the estimates provided by the baseline methods over the $10,000$ independent iterations and summarise them in Figure 6.

We can make two observations from the plots in Figure 6. First, when the angle θ between the two vectors is fixed, the squared norm of the vectors does not affect the convergence of all five methods, as each horizontal row displays a similar trend. Second, there is varying performance among the inner product estimates obtained via different root-finding algorithms when vectors are either close to each other ($\theta = \frac{\pi}{12}$) or far apart from each other ($\theta = \frac{11\pi}{12}$).

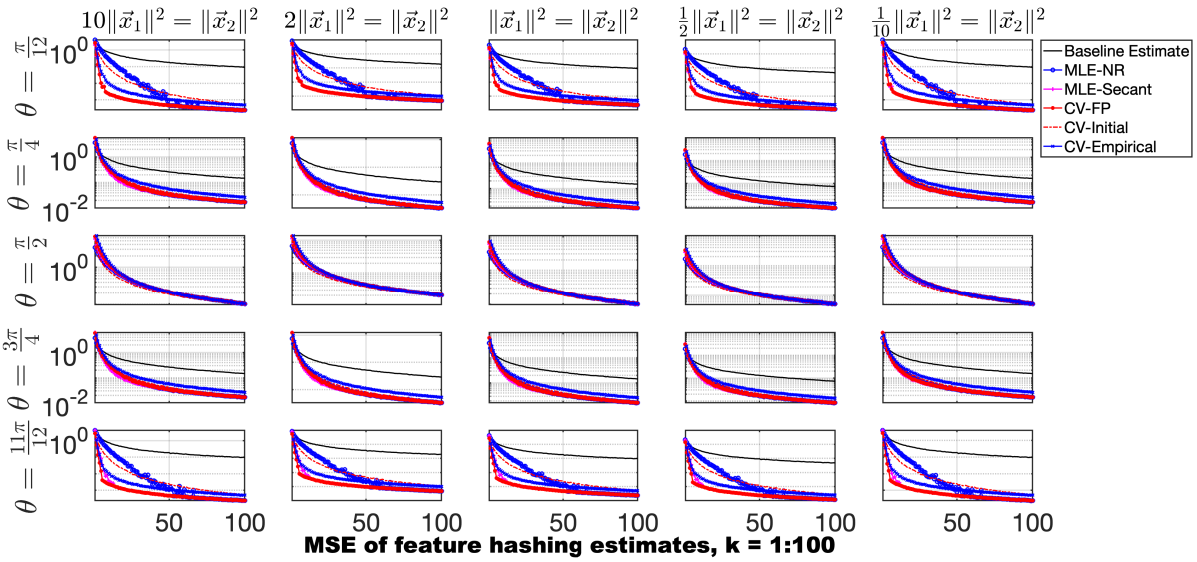

Figure 6: Plot of mean squared error (MSE) of respective estimates of inner product between $\vec{x}_1, \vec{x}_2$ as the squared vector norm and angle θ between the vectors vary for feature hashing. A lower value is better.

Further, the plots depicted in Figure 7 examine the case where $\|\vec{x}_1\|^2 = \|\vec{x}_2\|^2$, and $\theta = \frac{\pi}{12}$ and summarizes the MSE of five baseline methods. From Figure 7, it is clear that `MLE-NR` has inferior performance (excluding the baseline random projection estimate) until the number of observations k surpasses a certain value (≈ 70); after that, it has similar performance to `CV-FP` and `MLE-Secant`. For the smaller value of k, the MSE of `CV-Emp` is lower than `CV-Init`. However, the MSE of both algorithms converges as k increases. `MLE-Secant` and `CV-FP` have the same performance, which is superior to the other algorithms.

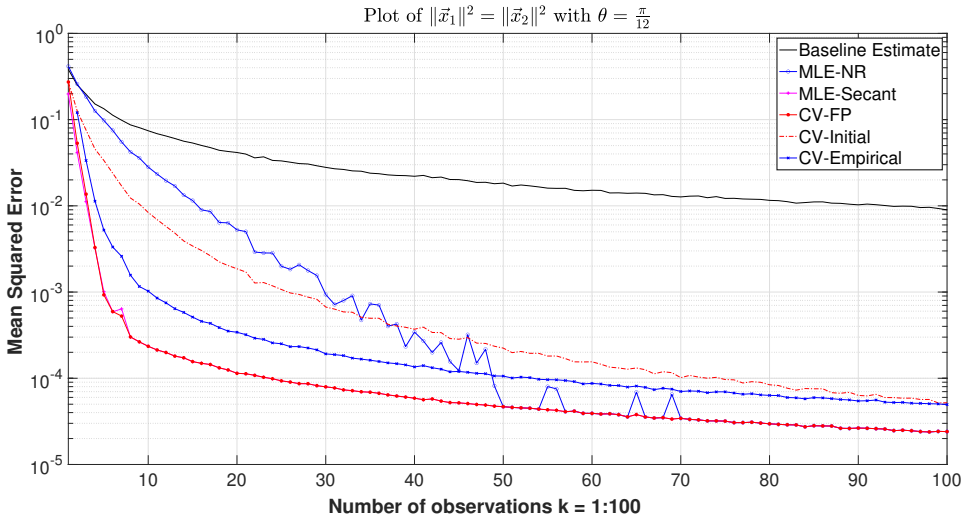

Figure 7: MSE plot when $\|\vec{x}_1\|^2 = \|\vec{x}_2\|^2$ with $\theta = \frac{\pi}{12}$ for feature hashing. A lower value is better.

(ii) **Proportion of outliers in** $10,000$ **estimates:** A smaller proportion of outliers suggests better performance. We generate the boxplots of the estimates of the inner product obtained using baseline methods over $10,000$ runs and present them in Figure 8 for $k = 10$ and $k = 20$. We summarise the corresponding proportion of outliers and a fraction of cubic equations with three real roots in Table 1.

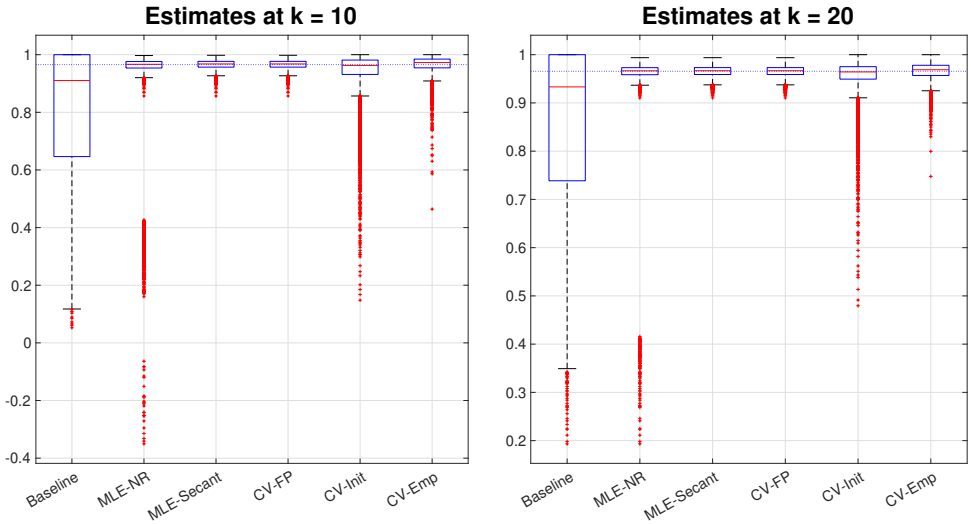

Figure 8: Boxplots of estimated inner products via the baseline feature hashing estimate (`FH`), Newton Raphson (`MLE-NR`), Secant method (`MLE-Secant`), our algorithm (`CV-FP`), CV using initial estimate (`CV-Init`), and CV using empirical values of variances and covariances (`CV-Emp`). The blue horizontal line denotes the true inner product between the vectors $\vec{x}_1, \vec{x}_2$. A narrower boxplot with few outliers is better.

Proportion of Outliers / Cubic with three real roots		
Method	$k = 10$	$k = 20$
MLE-NR	0.1097	0.0871
MLE-Secant	0.0670	0.0758
CV-FP	0.0670	0.0759
CV-Init	0.1156	0.1959
CV-Emp	0.1020	0.0736
Cubic with three real roots	0.0026	0

Table 1: Proportion of outliers in the first five lines of the table. Proportion of cubics with three real roots for the five methods at $k = 10$ and $k = 20$.

From Figure 8, it is evident that the interquartile range of estimates using all five algorithms is considerably smaller compared to the baseline feature hashing estimate, and the median closely approximates the true inner product between the vectors $\langle \vec{x}_1, \vec{x}_2 \rangle$. Table 1 highlights that the proportion of outliers for both `MLE-NR` and `CV-Emp` decreases from $k = 10$ to $k = 20$, though `MLE-NR` exhibits outliers further from the true value compared to `CV-Emp` and `CV-Init` (as depicted in Figure 8). Furthermore, at $k = 10$, there are 26 out of $10,000$ cubic equations with three real roots, whereas at $k = 20$, there are none, indicating that the poor performance of NR at small k is not due to the fact that there are multiple roots. CV-FP and MLE-Secant have a narrower interquartile range and a lower proportion of outliers than the others, resulting in their superiority over other methods.

(iii) **Number of updates required by each method until convergence to the estimate at respective sketch size (k):** A smaller number of updates implies that the method converges faster to the root. As we repeat the experiments $10,000$ times, each time we record the number of updates `MLE-NR`, `MLE-Secant`, and `CV-FP` methods took to coverage to the root (i.e., inner product estimate) at different values of the sketch size (k). We generate the boxplots corresponding to the number of updates and summarise it in Figure 9.

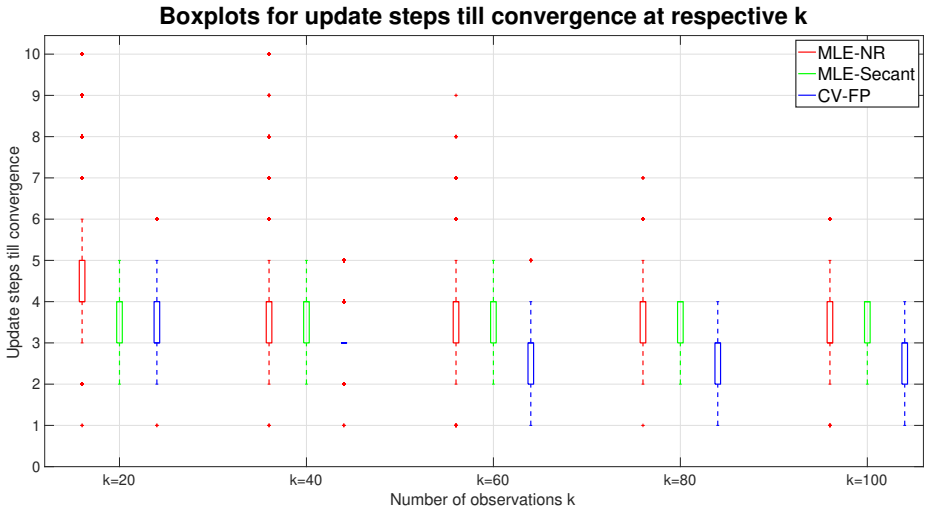

Figure 9: Boxplots of update steps for `MLE-NR`, `MLE-Secant`, and `CV-FP` at $k = \{20, 40, 60, 80, 100\}$ for feature hashing. A lower and narrower boxplot is better.

From Figure 9, it is clear that `CV-FP` takes fewer update steps to converge, and this holds generally throughout the different vector squared norms and varying angles θ. These experimental results indicate that `CV-FP` empirically achieves faster convergence than other baseline methods.

We also give a table of time taken for convergence to show that fewer number of update steps also implies faster time in Table 2.

C.3 RANDOM PROJECTION: FASTER CONVERGENCE AND BETTER STABILITY FOR THE MLE

Suppose $\mathbf{X}_{n\times p}$ is a data matrix, and $\mathbf{R}_{p\times k}$ is matrix with entries $r_{ij} \sim \mathcal{N}(0, 1)$. Consider the matrix $\mathbf{V}_{n\times k} = \mathbf{XR}$. For any two rows $\vec{v}_i, \vec{v}_j$ in the matrix, the distribution of the tuple $(v_{is}, v_{js}), 1 \le s \le k$ is bivariate normal, with

$$\begin{pmatrix} v_{is} \\ v_{js} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \|\vec{x}_i\|^2 & \langle \vec{x}_i, \vec{x}_j \rangle \\ \langle \vec{x}_i, \vec{x}_j \rangle & \|\vec{x}_j\|^2 \end{pmatrix} \right).$$

Li et al. (2006) used the fact that if the original lengths of the vectors $\vec{x}_1, \ldots, \vec{x}_n$ were known (e.g. normalized to the length of 1), then an MLE could be constructed to estimate $\langle \vec{x}_i, \vec{x}_j \rangle$. The MLE estimate of the inner product

	$\frac{\text{CV-FP}}{\text{Newton Raphson}}$	$\frac{\text{CV-FP}}{\text{Secant method}}$
10	0.905 ± 0.354	0.968 ± 0.414
20	0.882 ± 0.340	0.841 ± 0.295
30	0.910 ± 0.362	0.980 ± 0.364
40	0.918 ± 0.423	0.899 ± 0.416
50	0.877 ± 0.452	0.821 ± 0.432
60	0.961 ± 0.420	0.875 ± 0.392
70	0.945 ± 0.437	0.867 ± 0.416
80	0.956 ± 0.474	0.826 ± 0.392
90	0.890 ± 0.419	0.807 ± 0.340
100	0.892 ± 0.494	0.813 ± 0.448

Table 2: Ratios of time taken with 2 standard deviations over 10000 iterations with $\|\vec{x}_1\|^2 = \|\vec{x}_2\|^2, \theta = \frac{\pi}{12}$, for $k \in \{10, \ldots, 100\}$. A ratio less than 1 means that the method on the numerator is faster than the method on the denominator.

is the real root of the following cubic equation Li et al. (2006):

$$\lambda^3 - \left(\frac{\sum_{s=1}^{k} v_{is}v_{js}}{k}\right)\lambda^2 + \left(\|\vec{x}_i\|^2\left(\frac{\sum_{s=1}^{k} v_{js}^2}{k}\right) + \|\vec{x}_j\|^2\left(\frac{\sum_{s=1}^{k} v_{is}^2}{k}\right) - \|\vec{x}_i\|^2\|\vec{x}_j\|^2\right)\lambda - \left(\frac{\sum_{s=1}^{k} v_{is}v_{js}}{k}\right)\|\vec{x}_i\|^2\|\vec{x}_j\|^2 = 0, \tag{35}$$

and we note that Equation (35) is identical to Equation (32). Hence we follow the same steps to get the update step similar to feature hashing.

We adopt a similar experimental setup used for the feature hashing experiments in Section C.2.

We run simulations on vector pairs $\vec{x}_1, \vec{x}_2 \in \mathbb{R}^{100,000}$ with varying squared norms and angles between them over 10,000 iterations. We look at the ratios $r \in \{0.1, 0.5, 1, 2, 10\}$ where $\|\vec{x}_1\|^2 = r\|\vec{x}_2\|^2$, and angles $\theta \in \{\frac{\pi}{12}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \frac{11\pi}{12}\}$. For each iteration, we compute $V = XR$ and estimate the inner product as $\langle \vec{x}_i, \vec{x}_j \rangle \approx \frac{\sum s=1^k v_{is}v_{js}}{k}$ for $1 \leq k \leq 100$, referring to this estimate as the baseline estimate. We compute the estimates of the inner product using: a) MLE (Li et al., 2006) (Equation (35)) via Newton Raphson (`MLE-NR`) b) MLE (Equation (35)) via the Secant method (`MLE-Secant`), c) CV where the initial estimate $\frac{\sum_{s=1}^{k} v_{is}v_{js}}{k}$ is used as part of $\hat{c}_i$ (`CV-Init`), d) CV where the empirical covariances and variances are found for $\hat{c}_i$ (`CV-Emp`), and e) our CVE (Algorithm 1) (`CV-FP`). We record the inner product estimate provided by the above-mentioned methods and the number of updates each algorithm takes to converge the estimate in each iteration. Additionally, we note that cubic equations possess three real roots within the 10,000 iterations. We compare the performance of these methods using the following metrics:

(i) **Mean Square Error (MSE) w.r.t. k over $10,000$ iterations across various combinations of r and θ:** We compute the MSE of the estimates generated by baseline methods across various values of sketch size (k) over $10,000$ iterations for different combinations of r and θ. The MSE results are summarised in Figure 10.

From Figure 10, we can make two observations. First, for a fixed angle θ between the two vectors, the squared norm of the vectors does not affect the convergence of the baseline methods, since each horizontal row shows roughly the same trend. Second, there is varying performance between the inner product estimates via different root-finding algorithms when vectors are close to each other ($\theta = \frac{\pi}{12}$) or far away from each other ($\theta = \frac{11}{12}$).

The subsequent plots in Figure 11 depict the MSE of the baseline methods for $\|\vec{x}_1\|^2 = \|\vec{x}_2\|^2$ and $\theta = \frac{\pi}{12}$. It is evident from Figure 11 that `MLE-NR` exhibits the poorest performance (apart from the baseline random projection estimate) until the number of observations k surpasses a certain threshold (approximately 50), after which it demonstrates similar performance to `CV-FP` and `MLE-Secant`. At lower values of k, `CV-Emp` has a slightly higher MSE compared to `MLE-Secant` and `CV-FP`, while at higher values of k their MSE converges. The `MLE-Secant` and `CV-FP` exhibit nearly identical performance and outperform other algorithms across the entire range of sketch sizes (k).

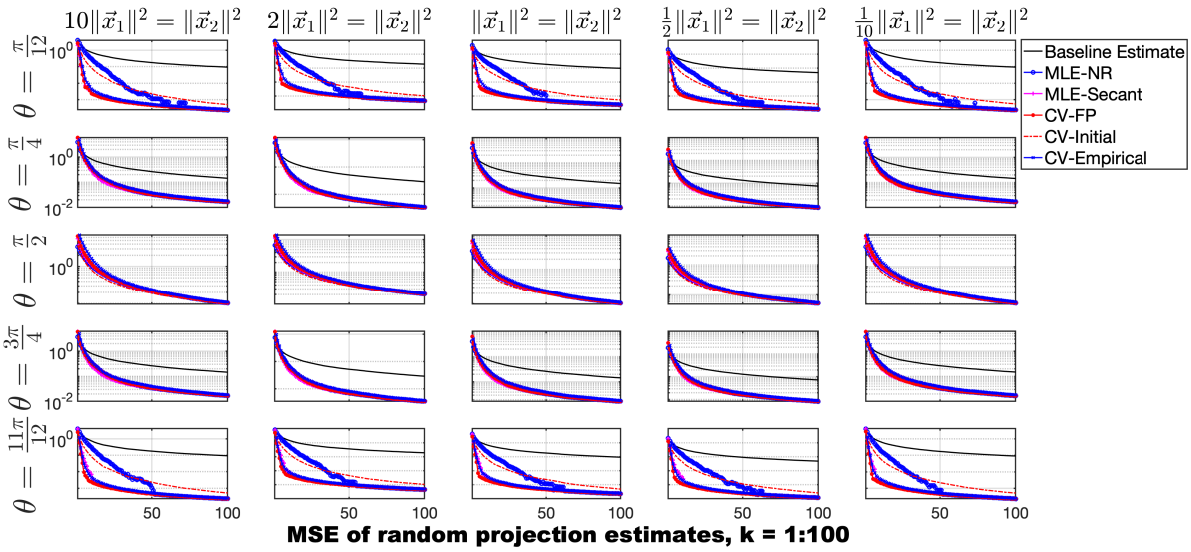

Figure 10: Plot of mean squared error (MSE) of respective estimates of inner product between $\vec{x}_1, \vec{x}_2$ as the squared vector norm and angle θ between the vectors vary for random projections. A lower value is better.

(ii) **Proportion of outliers in** $10,000$ **estimates:** For each baseline method, we construct boxplots using the $10,000$ estimates obtained over the 10,000 iterations. The boxplots for $k = 10$ and $k = 20$ are depicted in Figure 12. Furthermore, we provide a summary of the corresponding proportion of outliers in estimates and cubic equations with three real roots over 10,000 iterations in Table 3.

Proportion of Outliers / Cubic with three real roots		
Method	k = 10	k = 20
MLE-NR	0.1079	0.0861
MLE-Secant	0.0599	0.0737
CV-FP	0.0599	0.0737
CV-Init	0.1080	0.1936
CV-Emp	0.1057	0.0925
Cubic with three real roots	0.0021	0

Table 3: Proportion of outliers in the first five lines of the table. Proportion of cubics with three real roots for the five methods at $k = 10$ and $k = 20$.

From Figure 12, it is evident that the interquartile range of the estimates obtained using all five baseline methods is much smaller compared to the baseline random projection estimate, and the median is close to the true inner product between the two vectors $\langle \vec{x}_1, \vec{x}_2 \rangle$. From Table 3 we can see that the proportion of outliers for `MLE-NR` and `CV-Emp` decrease from $k = 10$ to $k = 20$, however `MLE-NR` has outliers farther away from the true value as compared to `CV-Emp` and `CV-Init` (Figure 12). Moreover, from Table 3, we can see that at $k = 10$, there are only 21 out of 10,000 cubics with three real roots, and at $k = 20$, there are 0 cubics with three real roots, implying that the poor performance of NR at small k cannot solely be attributed to the presence of cubics with multiple real roots. Finally, it is apparent from both Figure 12 and Table 3 that both `CV-FP` and `MLE-Secant` have a smaller interquartile range and a lower proportion of outliers compared to the other baseline methods, thus demonstrating their superiority over the other methods.

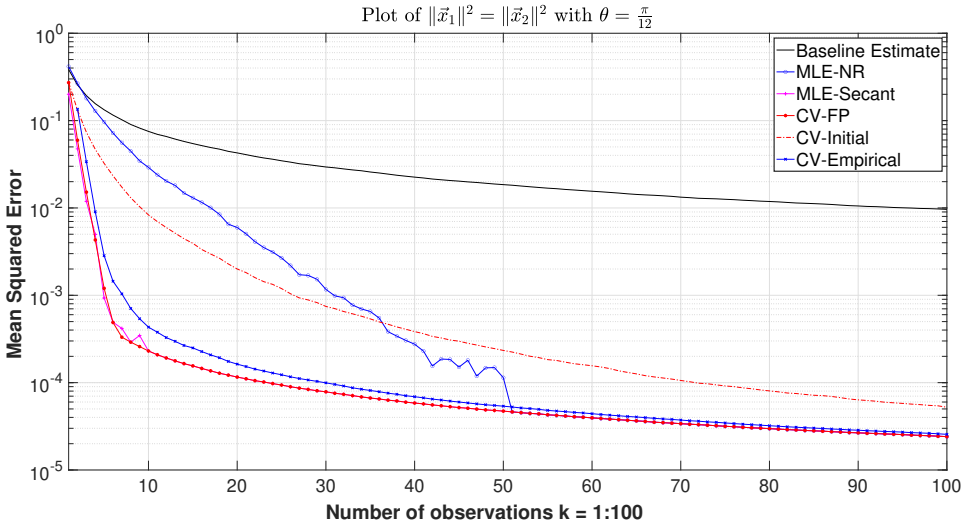

Figure 11: MSE plot when $\|\vec{x}_1\|^2 = \|\vec{x}_2\|^2$ with $\theta = \frac{\pi}{12}$ for random projections. A lower value is better.

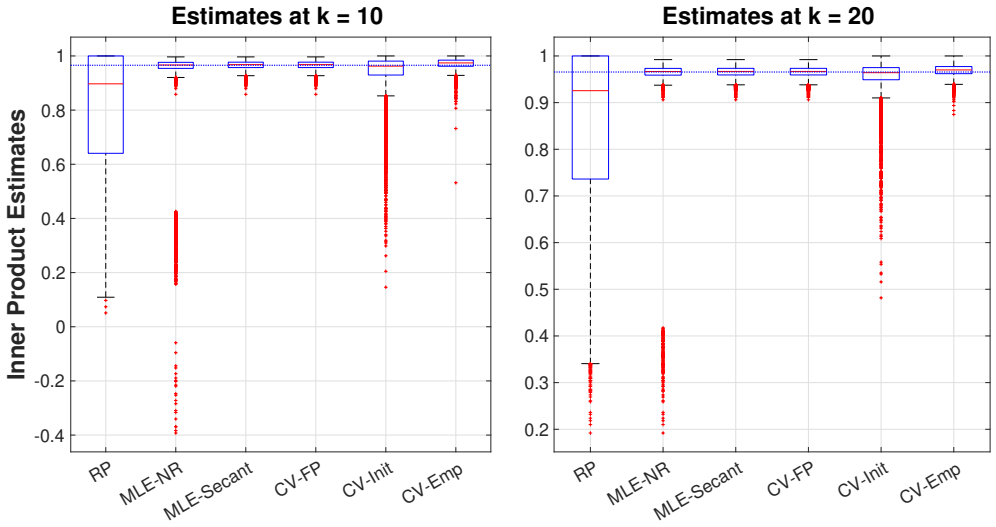

Figure 12: Boxplots of estimated inner products via the baseline random projection estimate (`RP`), Newton Raphson (`MLE-NR`), Secant method (`MLE-Secant`), our algorithm (`CV-FP`), CV using initial estimate (`CV-Init`), and CV using empirical values of variances and covariances (`CV-Emp`). The blue horizontal line denotes the true inner product between the vectors $\vec{x}_1, \vec{x}_2$. A narrower boxplot is better.

(iii) **Number of updates required by each method until convergence to the estimate at respective sketch size (k):** Lastly, we generate the boxplot corresponding to the number of updates each method took to converge to the inner product estimate in each of the $10,000$ iterations across various values of the sketch size (k). These results are summarized in Figure 13.

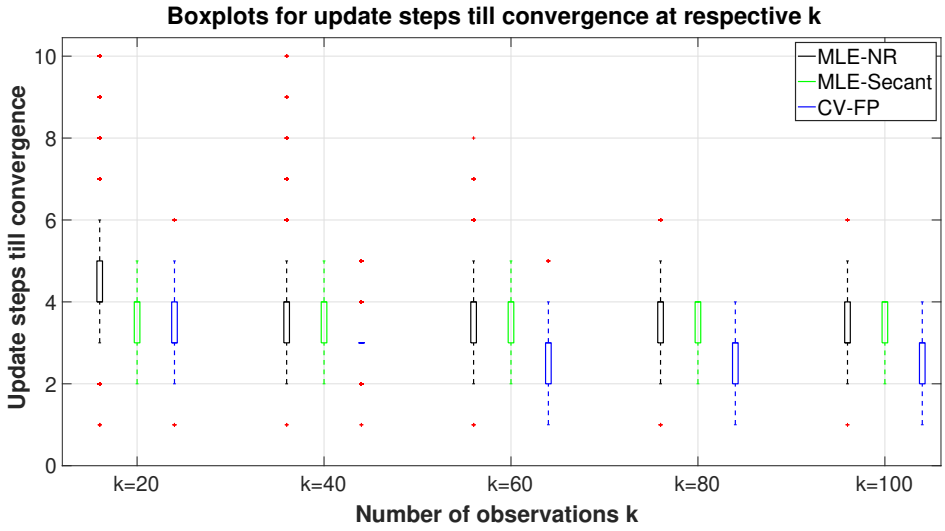

Figure 13: Boxplots of update steps for `MLE-NR`, `MLE-Secant`, and `CV-FP` at $k = \{20, 40, 60, 80, 100\}$ for random projections. A lower and narrower boxplot is better.

From Figure 13, it is evident that `CV-FP` requires fewer updates to converge to the estimate compared to `MLE-NR` and `MLE-Secant`. This trend holds generally throughout the different vector squared norms and varying angles θ, suggesting that empirically `CV-FP` achieves faster convergence to the estimate compared to the other methods.

C.4 IMPLICATION OF OUR EXPERIMENTS

We now summarize the implications of our experiments.

MLEs and CVEs are used to improve estimates when information about the parameters are known. However, most papers involving MLEs and CVEs focus on theoretical work (proving variance bounds) and experiments (comparing against benchmarked algorithms), but do not explicitly mention how the respective estimates from the CVE and MLE should be found, as the assumption is that a user can implement them on their own.

However, our plots show **there is a great difference** in the methods used, in terms of speed of convergence (Figure 9, Figure 13), and in terms of error (Figure 6, Figure 10), if a suboptimal implementation of `Newton Raphson` is used for the MLE, or computing $\hat{c}$ via `CV-Init` (or `CV-Emp`) via the CVE. This leads to possibly contradictory results based on the type of experiments conducted.

This leads to possibly contradictory results based on the type of experiments conducted.

For example, identifying vectors with high similarity is one goal of similarity search, where the angle θ between the two vectors are small. This is where a difference in methods used can lead to different performance (e.g. `Newton Raphson` does not have good performance with small k in Figure 9, Figure 13), but that does not mean an MLE is bad. Equivalently, outliers may be removed during experiments to show better results (our boxplots in Figure 9,Figure 13 show that `MLE-NR` has similar (good) performance with `MLE-Secant` and `CV-FP` ; but in a practical scenario the outliers are not known in advance.

Moreover, if sole error bars are used (based on computing the standard deviations of the estimates), this can lead to thinking that an MLE has "higher error", even though the outliers increase the standard deviations (Figure 8, Figure 12), and can be misleading. From Figure 8, Figure 12, there are more outliers in one direction for Newton Raphson. Using error bars implicitly assumes there is "equal variation above and below the mean", and may mask what is happening in practice.

Hence we suggest that Algorithm 1 (`CV-FP`) be used, which we expect will mitigate any reproducibility issues.

C.5 REMARKS ON CONVERGENCE RATE OF FIXED POINT ITERATION

Despite our experiments showing that `CV-FP` takes fewer update steps to converge compared to `MLE-NR` and `MLE-Secant`, `CV-FP` does not have quadratic convergence, and might not even have superlinear convergence.

Definition 1. *Suppose there is a sequence of real numbers* $\{x_i\}_{i=1}^n$ *such that* $\lim_{n\to\infty} = x$. *Further suppose*

$$\lim_{n\to\infty} \frac{|x_{n+1} - x|}{|x_n - x|^\alpha} = C < \infty. \tag{36}$$

If $\alpha = 1$, *the sequence converges linearly to* x. *If* $1 < \alpha < 2$, *the sequence converges superlinearly to* x. *If* $\alpha = 2$, *the sequence converges quadratically to* x.

The Newton Raphson algorithm has quadratic convergence, where $\alpha = 2$, and the Secant method has superlinear convergence, with $\alpha = \phi = 1.618...$ (Atkinson, 2008). To (empirically) check the convergence of an algorithm, where x_n are the updated values, and x the true value, we plot $\log(|x_{n+1} - x|)$ versus $\log(x_n - x)$ and record the gradient of the best fit line for our 10,000 iterations at $k = 100$ across all our vector pairs and angles (25,000 observations) in the bivariate Normal case for both feature hashing and random projections.

Equation (36) gives the motivation behind this, since

$$\begin{aligned} &|x_{n+1} - x| = C|x_n - x|^\alpha \\ \Rightarrow &\log(|x_{n+1} - x|) = \log(C) + \alpha \log(|x_n - x|). \end{aligned}$$

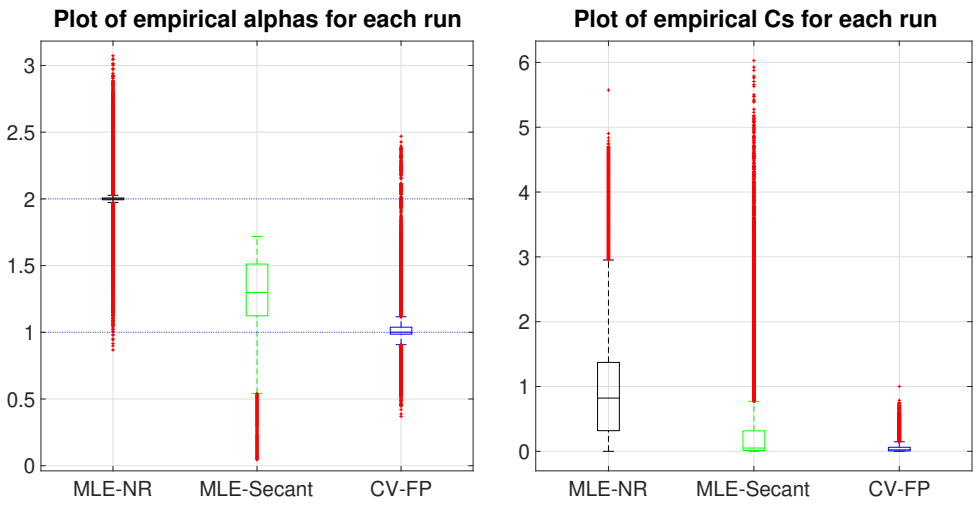

Figure 14: Boxplots of empirical α and C at $k = 100$ for `MLE-NR`, `MLE-Secant`, and `CV-FP` across all vector pairs at 25,000 observations for feature hashing.

The boxplots in Figures 14 and 15 show the empirical α and C for each run. The interquartile range of α for Newton Raphson is close to 2, but is less than ϕ for the Secant method. For `CV-FP`, the interquartile range is slightly above 1, and hence the convergence of `CV-FP` could either be linear or super-linear.

However, the empirical C for `CV-FP` is the lowest out of `MLE-NR` and `MLE-Secant`, which explains why there is faster convergence for `CV-FP` in the number of update steps required.

We leave this open for future work: to prove (or disprove) that the CVE has superlinear convergence, and that C is lower than both Newton Raphson and the Secant method **across all exponential families** satisfying the conditions in Theorem 4.

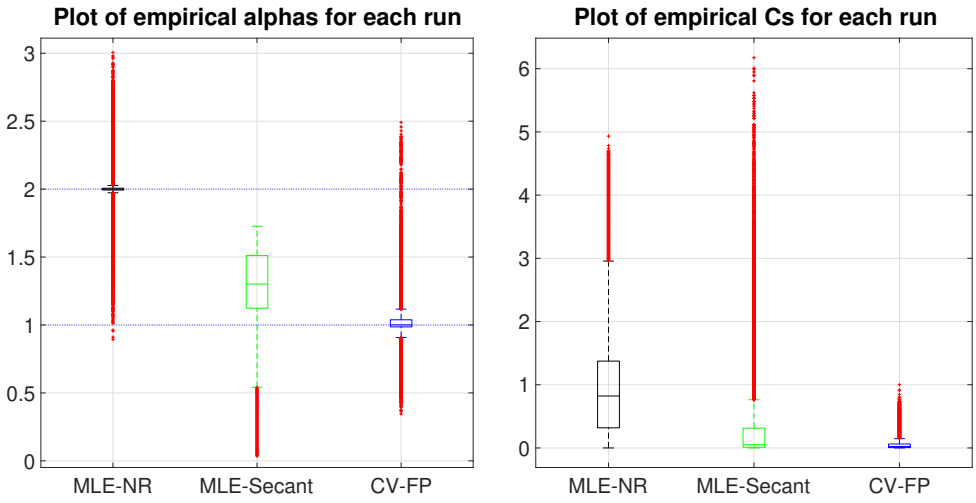

Figure 15: Boxplots of empirical α and C at $k = 100$ for `MLE-NR`, `MLE-Secant`, and `CV-FP` across all vector pairs at 25,000 observations for random projections.

D FURTHER DISCUSSION ON HUTCHINSON'S ESTIMATOR

This section of the appendix elaborates on our heuristic of how `CV-FP` can be used to find different estimators, if the conditions in Theorem 4 or Corollary 1, by either repeatedly iterating through the fixed point equations, or by finding a closed form solution analytically.

We demonstrate with Hutchinson's Estimator (Hutchinson, 1989) where we get an estimator, and analyze it directly.

Let $\vec{r}_i \in \mathbb{R}^d$ be distributed $\mathcal{N}(\vec{0}_d, \mathbf{I}_d)$. Let $y_i = \vec{r}_i^T \mathbf{M} \vec{r}_i \in \mathbb{R}$ for $1 \leq i \leq k$. For a symmetric $\mathbf{M}_{d\times d}$ matrix, the estimator $Y = \sum_{i=1}^k \frac{y_i}{k} = \sum_{i=1}^k \frac{\vec{r}_i^T \mathbf{M} \vec{r}_i}{k}$ estimates the trace of $\mathbf{M}$, with $\text{Var}(Y) = \frac{2\text{tr}(\mathbf{M}^2)}{k} = \frac{2\|\mathbf{M}\|_F^2}{k}$, where $\|\mathbf{M}\|_F^2$ is the Frobenius norm, defined by $\|\mathbf{M}\|_F^2 = \sum_{i,j} m_{ij}^2$.

Adams et al. (2018) gives the following CVE for $\text{tr}(\mathbf{M})$ when there is a known diagonal matrix $\mathbf{B}$,

$$Z = \sum_{i=1}^k \frac{\vec{r}_i^T \mathbf{M} \vec{r}_i}{k} + c\left(\sum_{i=1}^k \frac{\vec{r}_i^T \mathbf{B} \vec{r}_i}{k} - \text{tr}(\mathbf{B})\right) \tag{37}$$

Lemma 4.1 in Adams et al. (2018) gives the optimal $\hat{c}$ to be

$$\hat{c} = -\frac{\text{Cov}\left(\vec{r}^T \mathbf{M} \vec{r}, \vec{r}^T \mathbf{B} \vec{r}\right)}{\text{Var}\left(\vec{r}^T \mathbf{B} \vec{r}\right)} = -\frac{\text{tr}(\mathbf{MB})}{\text{tr}(\mathbf{B}^2)},$$

with a variance reduction of $\frac{2\text{tr}(\mathbf{MB})^2}{k\text{tr}(\mathbf{B}^2)}$. The empirical covariance $\text{Cov}\left(\vec{r}^T \mathbf{M} \vec{r}, \vec{r}^T \mathbf{B} \vec{r}\right)$ and variance $\text{Var}\left(\vec{r}^T \mathbf{B} \vec{r}\right)$ is computed from the observed data, since $\text{tr}(\mathbf{MB})$ is not known.

Given the work in Adams et al. (2018), a natural question would be to ask if a MLE could be found, which might yield a lower variance reduction. However, Y is distributed as a linear combination of χ^2 variables, and finding an MLE is non-trivial.

Heuristic 1: Find y_i in the CVE formulation which correspond to linearly independent sufficient statistics

Under the heuristic of y_i needing to be sufficient statistics (of parameters η_i), and since $\text{tr}(\mathbf{M}) = \sum_{i=1}^d m_{ii}$, we look for sufficient statistics of each m_{ii}. We decompose $\vec{r}^T \mathbf{M} \vec{r} = \sum_{j=1}^d r_j (\mathbf{M}\vec{r})_j$, where $(\mathbf{M}\vec{r})_j$ denotes the j^{th}

component of the vector $\mathbf{M}\vec{r}$, to get a CVE of the form

$$\begin{aligned} Z &= \sum_{j=1}^{d} \frac{\left(\sum_{i=1}^{k} r_{ij} \left(\mathbf{M}\vec{r}_i\right)_j\right)}{k} + \sum_{j=1}^{d} c_i \left(\frac{\sum_{i=1}^{k} r_{ij} \left(\mathbf{B}\vec{r}_i\right)_j}{k} - b_{ii}\right) \\ &= \tilde{y} + \sum_{j=1}^{d} c_i \left(\frac{\sum_{i=1}^{k} r_{ij} \left(\mathbf{B}\vec{r}_i\right)_j}{k} - b_{ii}\right), \end{aligned}$$

where each r_{is} is the s^{th} entry of $\vec{r}_i$.

Theorem 8. *For the CVE given by*

$$Z = \sum_{j=1}^{d} \left(\frac{\sum_{i=1}^{k} r_{ij} \left(\mathbf{M}\vec{r}_i\right)_j}{k}\right) + \sum_{j=1}^{d} c_i \left(\frac{\sum_{i=1}^{k} r_{ij} \left(\mathbf{B}\vec{r}_i\right)_j}{k} - b_{ii}\right),$$

the optimal coefficients $\hat{c}_i$ are given by $\hat{c}_i = -\frac{m_{ii}}{b_{ii}}$, with variance reduction given by $\frac{2}{k}\sum_{s=1}^{d} m_{ss}^2 \geq \frac{2\text{tr}(\mathbf{MB})^2}{k\text{tr}(\mathbf{B}^2)}$ when $\mathbf{B}$ is a diagonal matrix, where $b_{ii} \neq 0$.

Proof. The optimal coefficients $\hat{c}_i$ are given by

$$\begin{pmatrix} \text{Var}\left(r_1(\mathbf{B}\vec{r})_1\right) & \dots & \text{Cov}\left(r_1(\mathbf{B}\vec{r})_1, r_d(\mathbf{B}\vec{r})_d\right) \\ \vdots & \ddots & \vdots \\ \text{Cov}\left(r_d(\mathbf{B}\vec{r})_d, r_1(\mathbf{B}\vec{r})_1\right) & \dots & \text{Var}\left(r_d(\mathbf{B}\vec{r})_d\right) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_d \end{pmatrix} = -\begin{pmatrix} \text{Cov}\left(\tilde{y}, r_1(\mathbf{B}\vec{r})_1\right) \\ \vdots \\ \text{Cov}\left(\tilde{y}, r_d(\mathbf{B}\vec{r})_1\right) \end{pmatrix}, \tag{38}$$

noting that $\frac{1}{k^2}$ cancels out on both sides. Computing the entries of the variance-covariance matrix in Equation (38) gives

$$\begin{aligned} \text{Var}\left(r_s(\mathbf{B}\vec{r})_s\right) &= \|\vec{b}_s\|^2 + b_{ss}^2 \\ \text{Cov}\left(r_s(\mathbf{B}\vec{r})_s, r_t(\mathbf{B}\vec{r})_t\right) &= b_{st}b_{ts}, \end{aligned}$$

where $\vec{b}_s, \vec{b}_t$ denotes the $s^{\text{th}}, t^{\text{th}}$ row of $\mathbf{B}$. By linearity of covariances,

$$\text{Cov}\left(\tilde{y}, r_s(\mathbf{B}\vec{r})_s\right) = \sum_{i=1}^{d} \text{Cov}\left(r_i(\mathbf{M}\vec{r})_i, r_s(\mathbf{B}\vec{r})_s\right). \tag{39}$$

Equation (39) leads to

$$\begin{aligned} \text{Cov}\left(r_s(\mathbf{M}\vec{r})_s, r_s(\mathbf{B}\vec{r})_s\right) &= \left(\sum_{t=1}^{d} m_{st}b_{st}\right) + m_{ss}b_{ss} \\ \text{Cov}\left(r_s(\mathbf{M}\vec{r})_s, r_t(\mathbf{B}\vec{r})_t\right) &= m_{st}b_{ts}, \end{aligned}$$

and hence

$$\text{Cov}\left(\tilde{y}, r_s(\mathbf{B}\vec{r})_s\right) = 2m_{ss}b_{ss}.$$

As $\mathbf{B}$ is a diagonal matrix, the optimal coefficients $\hat{c}_i$ are now solved via

$$\begin{pmatrix} 2b_{11}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 2b_{dd}^2 \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_d \end{pmatrix} = -\begin{pmatrix} 2m_{11}b_{11} \\ \vdots \\ 2m_{dd}b_{dd} \end{pmatrix}$$

leading to $\hat{c}_i = -\frac{m_{ii}}{b_{ii}}$ and variance reduction given by

$$\begin{aligned} &\frac{1}{2}\begin{pmatrix} 2m_{11}b_{11} \\ \vdots \\ 2m_{dd}b_{dd} \end{pmatrix}^T \begin{pmatrix} \frac{1}{b_{11}^2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{b_{dd}^2} \end{pmatrix} \begin{pmatrix} 2m_{11}b_{11} \\ \vdots \\ 2m_{dd}b_{dd} \end{pmatrix} \\ &= \frac{2}{k}\sum_{s=1}^{d} m_{ss}^2. \end{aligned}$$

We now need to prove that the variance reduction is greater than the variance reduction given by Adams' CVE, and need to show that $2\sum_{s=1}^{d} m_{ss}^2 - 2\frac{\text{tr}(\mathbf{M})^2}{d} \geq 0$ and do so by the Cauchy-Schwartz Inequality. The Cauchy–Schwarz Inequality states that for any $m_s, b_s \in \mathbb{R}$,

$$\left(\sum_{s=1}^{d} m_s^2\right)\left(\sum_{s=1}^{d} b_s^2\right) \geq \left(\sum_{s=1}^{d} m_s b_s\right)^2. \tag{40}$$

Setting $m_s = m_{ss}$, and $b_s = 1, 1 \leq s \leq d$ in Equation (40) gives the identity

$$\begin{aligned}
&\left(\sum_{s=1}^{d} m_{ss}^2\right)\left(\sum_{s=1}^{d} 1^2\right) \geq \left(\sum_{s=1}^{d} m_{ss}\right)^2 \\
\Rightarrow &\sum_{s=1}^{d} m_{ss}^2 \geq \frac{1}{d}\left(\sum_{s=1}^{d} m_{ss}\right)^2 \\
\Rightarrow &\sum_{s=1}^{d} m_{ss}^2 \geq \frac{1}{d}\left(\text{tr}\left(M\right)\right)^2
\end{aligned}$$

with equality when m_{ss} are equal for all $1 \leq s \leq d$. □

From the proof of Theorem 8, the variance reduction from the CVE is independent of b, which implies $B = I$ should be used.

Heuristic 2: Apply Algorithm 1 given the CVE

While $\hat{c}_i$ is in terms of m_{ii}, we *consider the heuristic of taking the limit of the fixed point iteration, similar to Algorithm 1.* Suppose we have d control variates of the form

$$\frac{\sum_{i=1}^{k} r_{is}\left(\mathbf{M}\vec{r}_i\right)_s}{k} + \sum_{s=1}^{d} c_i\left(\frac{r_{is}\left(\mathbf{B}\vec{r}_i\right)_s}{k} - b_{ss}\right)$$

for each s. The optimal $\hat{c}_i$ are given by

$$\begin{pmatrix} 2b_{11}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 2b_{dd}^2 \end{pmatrix}\begin{pmatrix} c_1 \\ \vdots \\ c_d \end{pmatrix} = -\begin{pmatrix} \text{Cov}\left(r_s(\mathbf{M}\vec{r})_s, r_1(\mathbf{B}\vec{r})_1\right) \\ \vdots \\ \text{Cov}\left(r_s(\mathbf{M}\vec{r})_s, r_s(\mathbf{B}\vec{r})_s\right) \\ \vdots \\ \text{Cov}\left(r_s(\mathbf{M}\vec{r})_s, r_d(\mathbf{B}\vec{r})_d\right) \end{pmatrix} = -\begin{pmatrix} 0 \\ \vdots \\ 2m_{ss}b_{ss} \\ \vdots \\ 0 \end{pmatrix},$$

which implies the control variate coefficients $\hat{c}_i = 0, i \neq s$, giving the following estimate for m_{ss}

$$m_{ss} = \mathbb{E}\left(\sum_{i=1}^{k} \frac{r_{is}(\mathbf{M}\vec{r}_i)_s}{k} - \frac{m_{ss}}{b_{ss}}\left(\sum_{i=1}^{k} \frac{r_{is}(\mathbf{B}\vec{r}_i)_s}{k} - b_{ss}\right)\right). \tag{41}$$

Dropping the expectation in Equation (41), we get d fixed point iterations

$$f_{n+1} = \frac{\sum_{i=1}^{k} r_{is}(\mathbf{M}\vec{r}_i)_s}{k} - \frac{f_n}{b_{ss}}\left(\frac{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}{k} - b_{ss}\right), \quad 1 \leq s \leq d.$$

Assuming these fixed point iterations converge, we make f_n the subject to get

$$\widehat{m}_{ss} = \frac{b_{ss}\frac{\sum_{i=1}^{k} r_{is}(\mathbf{M}\vec{r}_i)_s}{k}}{\frac{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}{k}} = \frac{b_{ss}\sum_{i=1}^{k} r_{is}(\mathbf{M}\vec{r}_i)_s}{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}, \quad 1 \leq s \leq d.$$

and an estimate for $\operatorname{tr}(\mathbf{M})$ with k observations is

$$\widehat{\operatorname{tr}(\mathbf{M})} = \sum_{s=1}^{d} \frac{b_{ss} \sum_{i=1}^{k} r_{is}(\mathbf{M}\vec{r}_i)_s}{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}. \tag{42}$$

which matches (Bekas et al., 2007) when entries in $\vec{r}$ come from the Normal distribution. Moreover, the variance of the estimate of $\operatorname{tr}(\mathbf{M})$ is

$$\operatorname{Var}\left(\widehat{\operatorname{tr}(\mathbf{M})}\right) = \frac{2\|\mathbf{M}\|_F^2 - 2\sum_{s=1}^{d} m_{ss}^2}{k}, \tag{43}$$

with the intuition that error depends on the sum of the squares of the non-diagonal terms of $\mathbf{M}$.

While we do not know if the conditions for Theorem 4 have been satisfied, taking the limit of the fixed point iterations gives us the estimator of Bekas et al. (2007). Furthermore, this allows us to give some analysis on the approximate expectation and variance for this estimator, as Bekas et al. (2007) adopted a different approach.

Given that Equation (42) is exact, we do not need to check for convergence. However, we do need to verify that Equation (42) and Equation (43) are reasonable.

Proposition 1. *The expression* $\sum_{s=1}^{d} \frac{\sum_{i=1}^{k} b_{ss} r_{is}(\mathbf{M}\vec{r}_i)_s}{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}$ *is an estimator (possibly biased) of* $\operatorname{tr}(\mathbf{M})$ *with expectation (up to second order Taylor approximations) and variance (up to first-order Taylor approximations) given by*

$$\mathbb{E}\left(\sum_{s=1}^{d} \frac{\sum_{i=1}^{k} b_{ss} r_{is}(\mathbf{M}\vec{r}_i)_s}{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}\right) = \operatorname{tr}(\mathbf{M})$$

$$\operatorname{Var}\left(\sum_{s=1}^{d} \frac{\sum_{i=1}^{k} b_{ss} r_{is}(\mathbf{M}\vec{r}_i)_s}{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}\right) = \frac{2\|\mathbf{M}\|_F^2}{k} - \frac{2\sum_{s=1}^{d} m_{ss}^2}{k},$$

matching the expectation and variance in Theorem 8.

Proof. Let

$$b_{ss} \frac{\frac{\sum_{i=1}^{k} r_{is}(\mathbf{M}\vec{r}_i)_s}{k}}{\frac{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}{k}} = b_{ss} \frac{X_1^{(s)}}{X_2^{(s)}} = b_{ss} g(X_1^{(s)}, X_2^{(s)}). \tag{44}$$

To find an approximation for the expectation and variance of $g(X_1^{(s)}, X_2^{(s)})$, and covariance of $g(X_1^{(s)}, X_2^{(s)}), g(X_1^{(t)}, X_2^{(t)})$, we expand $g(X_1^{(s)}, X_2^{(s)})$ around $(\mathbb{E}\left(X_1^{(s)}\right), \mathbb{E}\left(X_2^{(s)}\right)) \equiv (m_{ss}, b_{ss})$ using the Taylor series approximation up to the second order terms. We write $\frac{\partial g^{(s)}}{\partial X_i^{(s)}}$ and $\frac{\partial^2 g^{(s)}}{\partial X_i^{(s)} \partial X_j^{(s)}}$ to mean the respective partial derivatives evaluated at $X_1^{(s)} = \mathbb{E}\left(X_1^{(s)}\right), X_2^{(s)} = \mathbb{E}\left(X_2^{(s)}\right)$. Hence we expand

$$\begin{aligned} g(X_1^{(s)}, X_2^{(s)}) = g\left(\mathbb{E}\left(X_1^{(s)}\right), \mathbb{E}\left(X_2^{(s)}\right)\right) + \frac{\partial g^{(s)}}{\partial X_1^{(s)}}(X_1^{(s)} - \mathbb{E}\left(X_1^{(s)}\right)) + \frac{\partial g^{(s)}}{\partial X_2^{(s)}}\left(X_2^{(s)} - \mathbb{E}\left(X_2^{(s)}\right)\right) \\ + \frac{1}{2}\left(\frac{\partial^2 g^{(s)}}{\partial {X_1^{(s)}}^2}\left(X_1 - \mathbb{E}\left(X_1^{(s)}\right)\right)^2 + \frac{\partial^2 g^{(s)}}{\partial {X_2^{(s)}}^2}\left(X_2^{(s)} - \mathbb{E}\left(X_2^{(s)}\right)\right)^2 \right. \\ \left. + 2\frac{\partial^2 g^{(s)}}{\partial X_1^{(s)} \partial X_2^{(s)}}\left(X_1^{(s)} - \mathbb{E}\left(X_1^{(s)}\right)\right)\left(X_2^{(s)} - \mathbb{E}\left(X_2^{(s)}\right)\right)\right) + R \end{aligned} \tag{45}$$

where R is a remainder term. The partial derivatives evaluated at $X_1^{(s)} = \mathbb{E}\left(X_1^{(s)}\right), X_2^{(s)} = \mathbb{E}\left(X_2^{(s)}\right)$ are

$$\frac{\partial g^{(s)}}{\partial X_1^{(s)}} = \frac{1}{\mathbb{E}\left(X_2^{(s)}\right)} = \frac{1}{b_{ss}}, \qquad \frac{\partial^2 g^{(s)}}{\partial {X_1^{(s)}}^2} = 0, \qquad \frac{\partial g^{(s)}}{\partial X_2^{(s)}} = -\frac{\mathbb{E}\left(X_1^{(s)}\right)}{\mathbb{E}\left(X_2^{(s)}\right)^2} = -\frac{m_{ss}}{b_{ss}^2}$$

$$\frac{\partial^2 g^{(s)}}{\partial X_1^{(s)} X_2^{(s)}} = \frac{\partial^2 g^{(s)}}{\partial X_2^{(s)} X_1^{(s)}} = -\frac{1}{\mathbb{E}\left(X_2^{(s)}\right)^2} = -\frac{1}{b_{ss}^2}, \qquad \frac{\partial^2 g^{(s)}}{\partial {X_2^{(s)}}^2} = \frac{2\mathbb{E}\left(X_1^{(s)}\right)}{\mathbb{E}\left(X_2^{(s)}\right)^3} = \frac{2m_{ss}}{b_{ss}^3}.$$

We use Equation (45) to compute expectations $\mathbb{E}\left(g(X_1^{(s)}, X_2^{(s)}\right)$ and get

$$\begin{aligned}\mathbb{E}\left(g(X_1^{(s)}, X_2^{(s)}\right) &= \frac{m_{ss}}{b_{ss}} + \frac{1}{2}\left(\frac{2m_{ss}}{b_{ss}^3}\mathrm{Var}\left(X_2\right) - \frac{2\mathrm{Cov}\left(X_1, X_2\right)}{b_{ss}^2}\right) + R \\ &= \frac{m_{ss}}{b_{ss}} + \frac{1}{2}\left(\frac{4m_{ss}b_{ss}^2}{kb_{ss}^3} - \frac{4m_{ss}b_{ss}}{kb_{ss}^2}\right) + R \\ &\approx \frac{m_{ss}}{b_{ss}}. \end{aligned} \tag{46}$$

Expanding out $\mathbb{E}\left(\sum_{s=1}^d \frac{b_{ss}\frac{\sum_{i=1}^k r_{is}(M\vec{r}_i)_s}{k}}{\frac{\sum_{i=1}^k r_{is}(B\vec{r}_i)_s}{k}}\right)$ via Equation (46) gives

$$\mathbb{E}\left(\sum_{s=1}^d \frac{b_{ss}\frac{\sum_{i=1}^k r_{is}(M\vec{r}_i)_s}{k}}{\frac{\sum_{i=1}^k r_{is}(B\vec{r}_i)_s}{k}}\right) = \mathbb{E}\left(\sum_{s=1}^d b_{ss}\mathbb{E}\left(g(X_1^{(s)}, X_2^{(s)}\right)\right) \approx \sum_{s=1}^d m_{ss} = \mathrm{tr}\left(M\right)$$

as desired. Similarly,

$$\begin{aligned}\mathbb{E}\left(g(X_1^{(s)}, X_2^{(s)})^2\right) &= \left(\frac{m_{ss}}{b_{ss}}\right)^2 + \frac{\mathrm{Var}\left(X_1\right)}{b_{ss}^2} + \frac{m_{ss}^2\mathrm{Var}\left(X_2\right)}{b_{ss}^4} - 2\frac{m_{ss}\mathrm{Cov}\left(X_1, X_2\right)}{b_{ss}^3} + R \\ &= \left(\frac{m_{ss}}{b_{ss}}\right)^2 + \frac{\|\vec{a}_s\|^2 + m_{ss}^2}{kb_{ss}^2} + \frac{2m_{ss}^2b_{ss}^2}{kb_{ss}^4} - \frac{4m_{ss}^2b_{ss}}{kb_{ss}^3} + R \\ &= \left(\frac{m_{ss}}{b_{ss}}\right)^2 + \frac{\|\vec{a}_s\|^2 + m_{ss}^2}{kb_{ss}^2} - \frac{2m_{ss}^2}{kb_{ss}^2} + R, \end{aligned} \tag{47}$$

and we compute $\mathrm{Var}\left(g(X_1^{(s)}, X_2^{(s)})\right)$ by Equations (46) and (46), giving

$$\begin{aligned}\mathrm{Var}\left(g(X_1^{(s)}, X_2^{(s)})\right) &= \mathbb{E}\left(g(X_1^{(s)}, X_2^{(s)})^2\right) - \mathbb{E}\left(g(X_1^{(s)}, X_2^{(s)})\right)^2 + R \\ &\approx \left(\frac{m_{ss}}{b_{ss}}\right)^2 + \frac{\|\vec{a}_s\|^2 + m_{ss}^2}{kb_{ss}^2} - \frac{2m_{ss}^2}{kb_{ss}^2} - \left(\frac{m_{ss}}{b_{ss}}\right)^2 \\ &= \frac{\|\vec{a}_s\|^2 + m_{ss}^2}{kb_{ss}^2} - \frac{2m_{ss}^2}{kb_{ss}^2}. \end{aligned} \tag{48}$$

We now consider the expectation of the product of $g(X_1^{(s)}, X_2^{(s)})g(X_1^{(t)}, X_2^{(t)})$ using first order terms in the Taylor expansion for any $1 \le s, t \le d$.

$$\begin{aligned}\mathbb{E}\left(g(X_1^{(s)}, X_2^{(s)})g(X_1^{(t)}, X_2^{(t)})\right) &= \frac{m_{ss}}{b_{ss}}\frac{m_{tt}}{b_{tt}} + \frac{\partial g^{(s)}}{\partial X_1^{(s)}}\frac{\partial g^{(t)}}{\partial X_1^{(t)}}\mathrm{Cov}\left(X_1^{(s)}, X_1^{(t)}\right) \\ &\quad + \frac{\partial g^{(s)}}{\partial X_1^{(s)}}\frac{\partial g^{(t)}}{\partial X_2^{(t)}}\mathrm{Cov}\left(X_1^{(s)}, X_2^{(t)}\right) + \frac{\partial g^{(s)}}{\partial X_2^{(s)}}\frac{\partial g^{(t)}}{\partial X_1^{(t)}}\mathrm{Cov}\left(X_2^{(s)}, X_1^{(t)}\right) \\ &\quad + \frac{\partial g^{(s)}}{\partial X_2^{(s)}}\frac{\partial g^{(t)}}{\partial X_2^{(t)}}\mathrm{Cov}\left(X_2^{(s)}, X_2^{(t)}\right) + R \\ &= \frac{m_{ss}}{b_{ss}}\frac{m_{tt}}{b_{tt}} + \frac{\mathrm{Cov}\left(X_1^{(s)}), X_1^{(t)}\right)}{b_{ss}b_{tt}} - \frac{\mathrm{Cov}\left(X_1^{(s)}, X_2^{(t)}\right)m_{tt}}{b_{ss}b_{tt}^2} \\ &\quad - \frac{\mathrm{Cov}\left(X_2^{(s)}, X_1^{(t)}\right)m_{ss}}{b_{ss}^2b_{tt}} + \frac{\mathrm{Cov}\left(X_2^{(s)}, X_2^{(t)}\right)m_{ss}m_{tt}}{b_{ss}^2b_{tt}^2} + R \\ &= \frac{m_{ss}}{b_{ss}}\frac{m_{tt}}{b_{tt}} + \frac{m_{st}m_{ts}}{kb_{ss}b_{tt}} - \frac{m_{st}b_{ts}m_{tt}}{kb_{ss}b_{tt}^2} - \frac{m_{ts}b_{st}m_{ss}}{kb_{ss}^2b_{tt}} + \frac{b_{st}b_{ts}m_{ss}m_{tt}}{kb_{ss}^2b_{tt}^2} + R \\ &\approx \frac{m_{ss}}{b_{ss}}\frac{m_{tt}}{b_{tt}} + \frac{m_{st}^2}{kb_{ss}b_{tt}}, \end{aligned} \tag{49}$$

where Equation (49) follows due to the fact that $\mathbf{B}$ was chosen to be diagonal, hence $b_{st} = 0, s \neq t$, and $\mathbf{M}$ is symmetric, so $m_{st} = m_{ts}$. This gives $\text{Cov}\left(g(X_1^{(s)}, X_2^{(s)}), g(X_1^{(t)}, X_2^{(t)})\right)$ to be

$$\text{Cov}\left(g(X_1^{(s)}, X_2^{(s)}), g(X_1^{(t)}, X_2^{(t)})\right) = \frac{m_{ss}}{b_{ss}}\frac{m_{tt}}{b_{tt}} + \frac{m_{st}^2}{kb_{ss}b_{tt}} - \frac{m_{ss}}{b_{ss}}\frac{m_{tt}}{b_{tt}} = \frac{m_{st}^2}{kb_{ss}b_{tt}}. \tag{50}$$

Combining Equations (48) and (50) finally gives

$$\begin{aligned}
\text{Var}\left(\sum_{s=1}^{d} \frac{b_{ss}\frac{\sum_{i=1}^{k} r_{is}(\mathbf{M}\vec{r}_i)_s}{k}}{\frac{\sum_{i=1}^{k} r_{is}(\mathbf{B}\vec{r}_i)_s}{k}}\right) &= \sum_{s=1}^{d} \text{Var}\left(b_{ss}g(X_1^{(s)}, X_2^{(s)})\right) \\
&\quad + \sum_{s,t}^{d} \text{Cov}\left(b_{ss}g(X_1^{(s)}, X_2^{(s)}), b_{tt}g(X_1^{(t)}, X_2^{(t)})\right) \\
&= \sum_{s=1}^{d}\left(\frac{\|\vec{a}_s\|^2 + m_{ss}^2}{k} - \frac{2m_{ss}^2}{k}\right) + \sum_{s,t}^{d}\frac{m_{st}^2}{k} + R \\
&\approx \frac{2\|\mathbf{M}\|_F^2}{k} - \frac{2m_{ss}^2}{k}.
\end{aligned}$$

□

The implications of Proposition 1 implies Equation (42) as an estimator for $\text{tr}(\mathbf{M})$. More precisely, this estimator from (Bekas et al., 2007) naturally appears as the "limit" of the optimal CVE. Equation (43) also complements the probability bounds in (Baston and Nakatsukasa, 2022) for each diagonal term of $\mathbf{M}$.

To conclude, treating the CVE as a fixed point iteration and finding its closed form can lead to more efficient estimators than by using CVE weights once, and then the estimators analyzed directly. With respect to our field in sketching algorithms, our heuristic can lead to re-examining past work on estimators to give alternate proofs for variance reduction (and using these variances in probability bounds), and future work on understanding new estimators along similar lines, e.g. applying our work to find better estimators for estimating Euclidean distances under composition of functions (Leroux and Rademacher, 2024), or for the Jaccard similarity under circulant permutations (Li and Li, 2022). In both these cases, computing expectations and covariances (hence CVE weights) are much easier than analytically deriving the MLE involved.