Gaussian Approximations in High Dimensional Estimation

Neeraja Sahasrabudhe
IITB
(joint work with Prof. V. S. Borkar and Raaz Dwivedi)

September 8, 2014

Curse of Dimensionality

- Advancement in data collection and storage have enabled collection of huge amounts of information.

Curse of Dimensionality

- Advancement in data collection and storage have enabled collection of huge amounts of information.
- However, huge data presents substantial challenges to existing data analysis tools.

Curse of Dimensionality

- Advancement in data collection and storage have enabled collection of huge amounts of information.
- However, huge data presents substantial challenges to existing data analysis tools.
- Existing algorithms scale poorly with increase in number of dimensions of the data.

Curse of Dimensionality

- Advancement in data collection and storage have enabled collection of huge amounts of information.
- However, huge data presents substantial challenges to existing data analysis tools.
- Existing algorithms scale poorly with increase in number of dimensions of the data.
- This motivates mapping of data from high dimensional space to a lower dimensional space in a manner that prevents certain features/structure of the data.

Several estimation techniques in current use assume validity of Gaussian approximations for estimation purposes.
These ensemble methods have proven to work very well for high-dimensional data even when the distributions involved are not necessarily Gaussian.

- Marginals in higher dimension can be approximated by marginals in lower dimension.

- Marginals in higher dimension can be approximated by marginals in lower dimension.
- Marginals in the lower dimensional space are approximately Gaussian.

Notation

- (X, Y) - a random vector in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with isotropic log concave density.

Notation

- (X, Y) - a random vector in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with isotropic log concave density.
- $(\Gamma X, \Gamma Y)$ - orthogonal projection (on first k_1 and first k_2 coordinates respectively) of random vectors X and Y.

Isotropic log-concave density

Isotropic log-concave density

- Definition
  A function $f : \mathbb{R}^n \longrightarrow [0, \infty)$ is log-concave if for all $x, y \in \mathbb{R}^n$ and $0 < \lambda < 1$,
  $$f(\lambda x + (1-\lambda)y) \geq f(x)^{\lambda} f(y)^{1-\lambda}.$$

Isotropic log-concave density

- Definition
  A function $f : \mathbb{R}^n \longrightarrow [0, \infty)$ is log-concave if for all $x, y \in \mathbb{R}^n$ and $0 < \lambda < 1$,
  $$f(\lambda x + (1-\lambda)y) \geq f(x)^{\lambda} f(y)^{1-\lambda}.$$

Isotropic log-concave density

- **Definition**
  A function $f : \mathbb{R}^n \longrightarrow [0, \infty)$ is log-concave if for all $x, y \in \mathbb{R}^n$ and $0 < \lambda < 1$,
  $$f(\lambda x + (1-\lambda)y) \geq f(x)^{\lambda} f(y)^{1-\lambda}.$$

- **Definition**
  We say that $f : \mathbb{R}^n \longrightarrow [0, \infty)$ is isotropic if it is the density function of some random variable with zero mean and identity covariance matrix. That is, f is isotropic when
  $$\int_{\mathbb{R}^n} f(x)dx = 1, \ \int_{\mathbb{R}^n} xf(x)dx = 0$$
  $$\text{and} \quad \int_{\mathbb{R}^n} < x, \theta >^2 f(x)dx = \|\theta\|^2; \ \ \forall\, \theta \in \mathbb{R}^n$$

Extend the random vectors ΓX and ΓY in $\mathbb{R}^{k_1}$ and $\mathbb{R}^{k_2}$ to random vectors in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ by adding appropriate number of zeroes respectively. By abuse of notation, we denote these new vectors in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ by ΓX and ΓY.

- $\mu \approx \mu_{emp}$ and $R \approx R_{emp}$

- $\mu \approx \mu_{emp}$ and $R \approx R_{emp}$
- $E[\Gamma Y|\Gamma X]$ is determined by a random vector having approximately $\mathcal{N}(\mu, R)$ distribution.

- $\mu \approx \mu_{emp}$ and $R \approx R_{emp}$
- $E[\Gamma Y|\Gamma X]$ is determined by a random vector having approximately $\mathcal{N}(\mu, R)$ distribution.
- $E[Y|X] \approx E[\Gamma Y|\Gamma X]$

Show that $(\Gamma X, \Gamma Y) \approx \mathcal{N}(\mu, R)$ with,

$$\mu = (\mu_1, \mu_2)$$

and

$$R = \left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right)$$

a square matrix of size $n_1 + n_2$. Then,

$$E[\Gamma Y | \Gamma X] \approx -\mu_2 + R_{21} R_{11}^{-1} (\Gamma X + \mu_1)$$

where $R_{11} \in \mathbb{R}^{n_1 \times n_1}$

Low Dimensional Projections with Gaussian Densities

Theorem (Eldan and Klartag(2007))

Let $1 \leq l \leq n$ be an integer and let $K \subset \mathbb{R}^n$ be a convex body (compact convex set with non-empty interior). Let X be a random vector that is distributed uniformly in K, and suppose that X has zero mean and identity covariance matrix. Assume that $l \leq cn^k$. Then there exists a subset $\mathcal{E} \subset G_{n,l}$ with $\sigma_n(\mathcal{E}) \geq 1 - e^{-cn^{0.9}}$ such that for any $E \in \mathcal{E}$,

$$\sup_{A \subseteq E} |P\{Proj_E(X) \in A\} - \int_A \phi_E^l(x)dx| \leq \frac{1}{n^k},$$

where the supremum runs over all measurable sets $A \subset E$. Here $\phi_E^l(x)$ is the standard l-dimensional Gaussian density in E and $c, k > 0$ are universal constants.

Low Dimensional Projections with Gaussian Densities

Theorem (Klartag(2008))

Let X be an isotropic random vector in $\mathbb{R}^n$ with a log-concave density. Let $1 \leq l \leq n^{c_1}$ be an integer. Then there exists a subset $\mathcal{E} \subset G_{n,l}$ with $\sigma_{n,l}(\mathcal{E}) \geq 1 - Ce^{-n^{c_2}}$ such that for any $E \in \mathcal{E}$, the following holds. Denote by f_E the density of the random vector $Proj_E(X)$, then for all $x \in E$ with $\|x\| \leq n^{c_4}$,

$$\left| \frac{f_E(x)}{\phi_E^l(x)} - 1 \right| \leq \frac{C}{n^{c_3}},$$

Here $C, c_1, c_2, c_3, c_4 > 0$ are universal constants.

Low Dimensional Projections with Gaussian Densities

- Let G denote the product Grassmanian of all subspaces $S_1 \times S_2$ of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ where for $i = 1, 2$, $S_i \subset \mathbb{R}^{n_i}$ and $\dim(S_i) = k_i$.

Low Dimensional Projections with Gaussian Densities

- Let G denote the product Grassmanian of all subspaces $S_1 \times S_2$ of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ where for $i = 1, 2$, $S_i \subset \mathbb{R}^{n_i}$ and $\dim(S_i) = k_i$.
- Let σ denote the unique rotationally invariant probability measure on G.

Low Dimensional Projections with Gaussian Densities

- Let G denote the product Grassmanian of all subspaces $S_1 \times S_2$ of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ where for $i = 1, 2$, $S_i \subset \mathbb{R}^{n_i}$ and $\dim(S_i) = k_i$.
- Let σ denote the unique rotationally invariant probability measure on G.
- Then there exists $\mathcal{E} \subset G$ with $\sigma(\mathcal{E}) \geq 1 - e^{-(n_1+n_2)^{c_2}}$ such that for all $(x, y) \in E$ with $\|(x, y)\|_2 \leqslant (n_1 + n_2)^{c_4}$

$$\left| \frac{f_E(x, y)}{\phi_E(x, y)} - 1 \right| \leq \frac{C}{(n_1 + n_2)^{c_3}},$$

Low Dimensional Projections with Gaussian Densities

- Let G denote the product Grassmanian of all subspaces $S_1 \times S_2$ of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ where for $i = 1, 2$, $S_i \subset \mathbb{R}^{n_i}$ and $\dim(S_i) = k_i$.
- Let σ denote the unique rotationally invariant probability measure on G.
- Then there exists $\mathcal{E} \subset G$ with $\sigma(\mathcal{E}) \geq 1 - e^{-(n_1+n_2)^{c_2}}$ such that for all $(x, y) \in E$ with $\|(x, y)\|_2 \leqslant (n_1 + n_2)^{c_4}$

$$\left| \frac{f_E(x, y)}{\phi_E(x, y)} - 1 \right| \leq \frac{C}{(n_1 + n_2)^{c_3}},$$

- So, $Proj_E(X, Y)$ is approximately Gaussian with high probability (i.e., $\geq 1 - e^{-(n_1+n_2)^{c_2}}$).

Remains to show: $E[Y|X] \approx E[\Gamma Y|\Gamma X]$

- *Step 1:* $E[\Gamma Y|\Gamma X] \approx E[Y|\Gamma X]$

Remains to show: $E[Y|X] \approx E[\Gamma Y|\Gamma X]$

- *Step 1:* $E[\Gamma Y|\Gamma X] \approx E[Y|\Gamma X]$
- *Step 2:* $E[Y|\Gamma X] \approx E[Y|X]$

Johnson-Lindenstrauss Lemma

Theorem (JL-lemma (1984))

For any $0 < \epsilon < 1$ and any integer n, let k be a positive integer such that

$$k \geq 8\frac{\ln n}{\epsilon^2}$$

Then for any set V of n points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ such that for all $u, v \in V$

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u+v\|^2$$

Key idea

Define a suitable probability distribution $\mathcal{F}$ on the set of all linear maps $\mathbb{R}^n \longrightarrow \mathbb{R}^k$. Then,

Lemma

Given $\epsilon > 0$, if $T : \mathbb{R}^n \longrightarrow \mathbb{R}^k$ is a random linear mapping drawn from the distribution $\mathcal{F}$, then for every vector $x \in \mathbb{R}^d$ we have

$$P\{(1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|\} \geq 1 - \frac{1}{n^2}$$

Klartag and Mendelson generalized the notion of Johnson-Lindenstrauss Lemma to a general set and tried to reduce the dependence of k on n, where n is the dimension of the original space and k is the dimension of the subspace for projection.

Definition

For a metric space (T, d) define

$$\gamma_\alpha(T, d) = \inf \sup_{t \in T} \sum_{s=0}^{\infty} 2^{s/\alpha} d(t, T_s),$$

where the infimum is taken with respect to all subsets $T_s \subset T$ with cardinality $|T_s| \geq 2^{2^s}$ and $|T_0| = 1$.

Theorem (Klartag and Mendelson(2005))

Let $G_{n,k}$ be the Grassmanian of k-dimesional subspaces of $\mathbb{R}^n$ with the unique rotation invariant probability measure on $G_{n,k}$ denoted by $\sigma_{n,k}$. Then given $\epsilon > 0$, for $k \geq C\gamma_2^2(\mathbb{R}^n, \|.\|_2)/\epsilon^2$, the following holds with probability larger than $1/2$, for $\Gamma = \sqrt{n}P$ where P is an orthogonal projection on a random k-dimensional subspace of $\mathbb{R}^n$ drawn from $G_{n,k}$ as per $\sigma_{n,k}$:

$$1 - \epsilon \leq \|\Gamma x\| \leq 1 + \epsilon.$$

Repeating this projection $O(n)$ times can boost the success probability to a desired constant, giving us the claimed randomized polynomial time algorithm. Specifically, after repeated independent projections, say $a \geq 1$ times, we can choose the best (i.e., one with the maximum norm) projection to get

$$P(1-\epsilon \leq \|\Gamma x\| \leq 1+\epsilon) > 1 - 1/2^a$$

.

Martingale Difference Sequence

$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots$ be a a filtration of $\sigma-$field $\mathcal{F}$ of a measure space $(\Omega, \mathcal{F}, \mathcal{P})$. A sequence $Y_1, Y_2, \dots$ of random variables form a martingale difference sequence if Y_k is $\mathcal{F}_k$-measurable and $E(Y_k|\mathcal{F}_{k-1}) = 0$ for each positive integer k.

- Given the random vector (X, Y), let $X = (X_1, \ldots, X_{n_1})$. For $1 \leq s \leq n_1$, let us denote by $\Gamma_s X$ the projection on first s coordinates.

- Given the random vector (X, Y), let $X = (X_1, \ldots, X_{n_1})$. For $1 \leq s \leq n_1$, let us denote by $\Gamma_s X$ the projection on first s coordinates.
- Again, by abuse of notation we denote the vector in $\mathbb{R}^{n_1}$ obtained by adding $n_1 - s$ zeroes at the end as $\Gamma_s X$.

- Given the random vector (X, Y), let $X = (X_1, \ldots, X_{n_1})$. For $1 \leq s \leq n_1$, let us denote by $\Gamma_s X$ the projection on first s coordinates.
- Again, by abuse of notation we denote the vector in $\mathbb{R}^{n_1}$ obtained by adding $n_1 - s$ zeroes at the end as $\Gamma_s X$.
- So, $E[Y|\Gamma_{n_1} X] = E[Y|X]$ and $E[Y|\Gamma_{k_1} X] = E[Y|\Gamma X]$.

- Given the random vector (X, Y), let $X = (X_1, \ldots, X_{n_1})$. For $1 \leq s \leq n_1$, let us denote by $\Gamma_s X$ the projection on first s coordinates.
- Again, by abuse of notation we denote the vector in $\mathbb{R}^{n_1}$ obtained by adding $n_1 - s$ zeroes at the end as $\Gamma_s X$.
- So, $E[Y|\Gamma_{n_1} X] = E[Y|X]$ and $E[Y|\Gamma_{k_1} X] = E[Y|\Gamma X]$.
- $E[Y|\Gamma_{s+1} X] - E[Y|\Gamma_s X]$ form a martingale difference family for $1 \leq s \leq n_1 - 1$.

Theorem (Hoeffding-Azuma inequality)

Let $\alpha_1, \alpha_2, \ldots$ be constants, and let $Y_1, Y_2 \ldots$ be a martingale difference sequence with $|Y_k| \leq \alpha_k$ for each k. Then for any $t \geq 0$,

$$\mathcal{P}\left(\sum Y_k \geq t\right) \leq 2e^{-t^2/2\sum \alpha_k^2}.$$

Let $\{Z_i\}$ be a martingale difference sequence. Define $S_n = \sum_{i=1}^{n} Z_i$.

- Lesigne and Volny(2001). If $\sup_i E[e^{|Z_i|}] < \infty$, then $\exists\ c > 0$ such that $P(S_n > n) \leq e^{-cn^{1/3}}$

Let $\{Z_i\}$ be a martingale difference sequence. Define $S_n = \sum_{i=1}^{n} Z_i$.

- Lesigne and Volny(2001). If $\sup_i E[e^{|Z_i|}] < \infty$, then $\exists\ c > 0$ such that $P(S_n > n) \leq e^{-cn^{1/3}}$
- Finite pth moments ($p \geq 2$): $P(S_n > n) \leq cn^{-p/2}$

Let $\{Z_i\}$ be a martingale difference sequence. Define $S_n = \sum_{i=1}^{n} Z_i$.

- Lesigne and Volny(2001). If $\sup_i E[e^{|Z_i|}] < \infty$, then $\exists\ c > 0$ such that $P(S_n > n) \leq e^{-cn^{1/3}}$
- Finite pth moments ($p \geq 2$): $P(S_n > n) \leq cn^{-p/2}$
- Y. Li (2003). If $Z_i \in \mathcal{L}^p$, $1 < p \leq 2, \|Z_i\| \leq M$ for all i, and let $x > 0$. Then

$$P(|S_n| > nx) \leq \frac{M^p}{x^p} b_p^p n^{1-p},$$

where $b_p = 18pq^{1/2}$ and q is such that $1/p + 1/q = 1$.

So we have,

- $P(\|E[Y|\Gamma X] - E[Y|X]\| > \epsilon) \leq e^{-c(n_1 - k_1 - 1)^{1/3}}$
  provided $\sup_i E[e^{\|E[Y|\Gamma_i X]\|}] < \infty$

So we have,

- $P(\|E[Y|\Gamma X] - E[Y|X]\| > \epsilon) \leq e^{-c(n_1-k_1-1)^{1/3}}$
  provided $\sup_i E[e^{\|E[Y|\Gamma_i X]\|}] < \infty$
- For $E[Y|\Gamma_{s+1}X] - E[Y|\Gamma_s X] \in \mathcal{L}^2$
  and $\|E[Y|\Gamma_{s+1}X] - E[Y|\Gamma_s X]\| \leq M$

$$P(S_n > n) \leq o(n^{-1}).$$

- Exponential finite moments:

$$P(\|E[\Gamma Y|\Gamma X] - E[Y|X]\| > \epsilon) \leq \frac{1}{2^a} + e^{-c(n_1 - k_1 - 1)^{1/3}}.$$

- Exponential finite moments:

$$P(\|E[\Gamma Y|\Gamma X] - E[Y|X]\| > \epsilon) \leq \frac{1}{2^a} + e^{-c(n_1-k_1-1)^{1/3}}.$$

- Bounded in $\mathcal{L}^p$:

$$P(\|E[\Gamma Y|\Gamma X] - E[Y|X]\| > \epsilon) \leq \frac{1}{2^a} + o(n^{-1/3}).$$

- Exponential finite moments:

$$P(\|E[\Gamma Y|\Gamma X] - E[Y|X]\| > \epsilon) \leq \frac{1}{2^a} + e^{-c(n_1-k_1-1)^{1/3}}.$$

- Bounded in $\mathcal{L}^p$:

$$P(\|E[\Gamma Y|\Gamma X] - E[Y|X]\| > \epsilon) \leq \frac{1}{2^a} + o(n^{-1/3}).$$

- Combine this with the fact that $E[\Gamma Y|\Gamma X]$ is Gaussian with probability larger than $1 - e^{-(n_1+n_2)^{c_2}}$.

Thank you!