Discussion of

Stein method in Bayesian computation

Nicolas Chopin

ENSAE, IPP (Institut Polytechnique de Paris)

Introduction

- For lack of time (and expertise), I will focus on control variates.

Introduction

- For lack of time (and expertise), I will focus on control variates.
- However, I will say a few words about the generality of Stein method near the end.

Control variates in a nutshell

To understand control variates, consider the following problem: we have IID pairs (X_n, Y_n), $n = 1, \ldots, N$, such that $\mathbb{E}(X_n) = 0$. To estimate $\alpha = \mathbb{E}(Y)$, we could use:

1. The empirical mean: $\bar{Y} = N^{-1}(Y_1 + \ldots + Y_N)$.

Control variates in a nutshell

To understand control variates, consider the following problem: we have IID pairs (X_n, Y_n), $n = 1, \ldots, N$, such that $\mathbb{E}(X_n) = 0$. To estimate $\alpha = \mathbb{E}(Y)$, we could use:

1. The empirical mean: $\bar{Y} = N^{-1}(Y_1 + \ldots + Y_N)$.
2. The OLS estimate corresponding to the following linear regression:

$$Y_n = \alpha + \beta X_n + \epsilon_n$$

where the ϵ_n are noise terms (zero mean).

Control variates in a nutshell

To understand control variates, consider the following problem: we have IID pairs (X_n, Y_n), $n = 1, \ldots, N$, such that $\mathbb{E}(X_n) = 0$. To estimate $\alpha = \mathbb{E}(Y)$, we could use:

1. The empirical mean: $\bar{Y} = N^{-1}(Y_1 + \ldots + Y_N)$.
2. The OLS estimate corresponding to the following linear regression:

$$Y_n = \alpha + \beta X_n + \epsilon_n$$

where the ϵ_n are noise terms (zero mean).

Control variates in a nutshell

To understand control variates, consider the following problem: we have IID pairs (X_n, Y_n), $n = 1, \ldots, N$, such that $\mathbb{E}(X_n) = 0$. To estimate $\alpha = \mathbb{E}(Y)$, we could use:

1. The empirical mean: $\bar{Y} = N^{-1}(Y_1 + \ldots + Y_N)$.
2. The OLS estimate corresponding to the following linear regression:

$$Y_n = \alpha + \beta X_n + \epsilon_n$$

   where the ϵ_n are noise terms (zero mean).

By construction, estimate 2 always outperforms estimate 1. By how much?

Control variates in a nutshell

To understand control variates, consider the following problem: we have IID pairs (X_n, Y_n), $n = 1, \ldots, N$, such that $\mathbb{E}(X_n) = 0$. To estimate $\alpha = \mathbb{E}(Y)$, we could use:

1. The empirical mean: $\bar{Y} = N^{-1}(Y_1 + \ldots + Y_N)$.
2. The OLS estimate corresponding to the following linear regression:
$$Y_n = \alpha + \beta X_n + \epsilon_n$$
where the ϵ_n are noise terms (zero mean).

By construction, estimate 2 always outperforms estimate 1. By how much?

Look at the R^2.

Generalisations

- replace the X_n by vectors of dimension p: multivariate regression. Note the $\mathcal{O}(p^3)$ complexity.

Generalisations

- replace the X_n by vectors of dimension p: multivariate regression. Note the $\mathcal{O}(p^3)$ complexity.
- Automatically choose certain components: Lasso.

Generalisations

- replace the X_n by vectors of dimension p: multivariate regression. Note the $\mathcal{O}(p^3)$ complexity.
- Automatically choose certain components: Lasso.
- Extension: non-parametric regression.

Application to Monte Carlo

Suppose you have any algorithm that generate random variables $\Theta_1, \ldots, \Theta_N$ according to e.g. a posterior distribution $\pi(d\theta)$. Ignore the fact they not be IID. Then:

1. Take $Y_n = \varphi(\Theta_n)$ for any $\varphi : \Theta \to \mathbb{R}$ of interest;

Application to Monte Carlo

Suppose you have any algorithm that generate random variables $\Theta_1, \ldots, \Theta_N$ according to e.g. a posterior distribution $\pi(d\theta)$. Ignore the fact they not be IID. Then:

1. Take $Y_n = \varphi(\Theta_n)$ for any $\varphi : \Theta \to \mathbb{R}$ of interest;
2. Find "by-products" X_n of the Θ_n's, which have expectation zero.

Application to Monte Carlo

Suppose you have any algorithm that generate random variables $\Theta_1, \ldots, \Theta_N$ according to e.g. a posterior distribution $\pi(d\theta)$. Ignore the fact they not be IID. Then:

1. Take $Y_n = \varphi(\Theta_n)$ for any $\varphi : \Theta \to \mathbb{R}$ of interest;
2. Find "by-products" X_n of the Θ_n's, which have expectation zero.
3. Linear regression.

Control variates: why nobody uses them?

For a long time, I thought CV were not popular mainly because it was a method that depends on the test function φ.

Control variates: why nobody uses them?

For a long time, I thought CV were not popular mainly because it was a method that depends on the test function φ.

However, this is a silly argument. The OLS estimate is:

$$\beta_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$$

and the only φ-dependent part is Y: pre-compute $(X^T X)^{-1} X^T$.

Control variates: why nobody uses them?

For a long time, I thought CV were not popular mainly because it was a method that depends on the test function φ.

However, this is a silly argument. The OLS estimate is:

$$\beta_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$$

and the only φ-dependent part is Y: pre-compute $(X^T X)^{-1} X^T$.

Remaining issues:

- how to construct control variates?

Control variates: why nobody uses them?

For a long time, I thought CV were not popular mainly because it was a method that depends on the test function φ.

However, this is a silly argument. The OLS estimate is:

$$\beta_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$$

and the only φ-dependent part is Y: pre-compute $(X^T X)^{-1} X^T$.

Remaining issues:

- how to construct control variates?
- complexity is $\mathcal{O}(p^3)$ if you take p covariates.

The curious link between control variates and invariant Markov processes

- One way to obtain CVs to use the infinitesimal generator of a process that leaves π invariant (e.g. Langevin in this talk).

The curious link between control variates and invariant Markov processes

- One way to obtain CVs to use the infinitesimal generator of a process that leaves π invariant (e.g. Langevin in this talk).
- Interestingly, you can also do the same with MCMC (discrete-time) kernels; in particular Gibbs samplers such that you are able to compute exactly $\mathbb{E}[\psi(X_t)|X_{t-1} = x]$ (Dellaportas and Kontoyiannis, 2012).

The curious link between control variates and invariant Markov processes

- One way to obtain CVs to use the infinitesimal generator of a process that leaves π invariant (e.g. Langevin in this talk).
- Interestingly, you can also do the same with MCMC (discrete-time) kernels; in particular Gibbs samplers such that you are able to compute exactly $\mathbb{E}[\psi(X_t)|X_{t-1} = x]$ (Dellaportas and Kontoyiannis, 2012).
- You can very well use one kernel to generate your random variables, and another kernel to construct control variates.

The curious link between control variates and invariant Markov processes

- One way to obtain CVs to use the infinitesimal generator of a process that leaves π invariant (e.g. Langevin in this talk).
- Interestingly, you can also do the same with MCMC (discrete-time) kernels; in particular Gibbs samplers such that you are able to compute exactly $\mathbb{E}[\psi(X_t)|X_{t-1} = x]$ (Dellaportas and Kontoyiannis, 2012).
- You can very well use one kernel to generate your random variables, and another kernel to construct control variates.
- Another interesting area of investigation: taking into account that your kernel does not simulate IID variables (e.g. Belomestny et al, 2020).

Conclusions

- you don't really *need* Stein method to construct control variates:

Conclusions

- you don't really *need* Stein method to construct control variates:
  1. you may use Markov process theory instead.

Conclusions

- you don't really *need* Stein method to construct control variates:
  1. you may use Markov process theory instead.
  2. the fact the class uniquely characterises the distribution does not seem to play any role.

Conclusions

- you don't really *need* Stein method to construct control variates:
    1. you may use Markov process theory instead.
    2. the fact the class uniquely characterises the distribution does not seem to play any role.
- Still the connection between CVs and Stein theory is neat, and the latter seems useful in many other areas, as the speaker showed us eloquently.

Conclusions

- you don't really *need* Stein method to construct control variates:
  1. you may use Markov process theory instead.
  2. the fact the class uniquely characterises the distribution does not seem to play any role.
- Still the connection between CVs and Stein theory is neat, and the latter seems useful in many other areas, as the speaker showed us eloquently.
- What about the $\mathcal{O}(n^2)$ complexity however?