A GRADIENT APPROXIMATION WITH IMPORTANCE SAMPLING FOR DIMENSION REDUCTION IN NATURAL EXPONENTIAL FAMILIES

Bastien Batardière, [1] Julien Chiquet [1], Joon Kwon [1]& Julien Stoehr [2]

[1] *MIA Paris-Saclay, Paris-Saclay University, AgroParisTech, INRAE, France, {prenom.nom}@inrae.fr*

[2] *Ceremade, Paris-Dauphine University, France, stoehr@ceremade.dauphine.fr*

Résumé. Les données de comptage en grande dimension sont difficiles à analyser telles quelles, et les approches fondées sur des modèles statistiques à variable latente restent efficaces et appropriées, tout en préservant l'explicabilité. Nous considérons plus particulièrement ici le cadre de modèles où les données discrètes sont guidées par une variable gaussienne latente décrivant la structure de dépendances des comptages dans un espaces de faible dimension, puis envoyées dans un espace de grande dimension via une distribution de Poisson ou Binomiale. Comme la loi de la variable latente conditionnement aux données reste inconnue, l'inférence variationnelle s'est révélée efficace pour inférer un tel modèle. Cependant, elle ne maximise qu'une borne inférieure de la vraisemblance et les estimateurs correspondant souffrent d'un manque de garanties théoriques. De plus, un grand nombre de paramètres variationnels est nécessaires. Dans ce travail en cours, nous utilisons l'échantillonnage préférentiel pour estimer les gradients de la log-vraisemblance. Nous contrôlons le biais de l'estimateur et nous appuyons sur des théorèmes d'optimisation pour assurer la convergence d'un schéma de gradient stochastique, s'adaptant facilement à un grand nombre d'échantillons.

Mots-clés. Données de comptage, optimisation, échantillonage préférentiel, descente de gradient, famille exponentielle naturelle.

Abstract. High-dimensional counting data is challenging to analyze as is, and approaches based on latent variable statistical models remain effective and appropriate, while preserving explainability. We particularly consider here the framework of models where discrete data are guided by a latent Gaussian variable describing the dependency structure of counts in a low-dimensional space, and then mapped into a high-dimensional space via a Poisson or Binomial distribution. Since the law of the latent variable conditioned on the data remains unknown, variational inference has proven effective in inferring such a model. However, it only maximizes a lower bound of the likelihood, and the corresponding estimators suffer from a lack of theoretical guarantees. Additionally, a large number of variational parameters are required. In this ongoing work, we use importance sampling to estimate the gradients of the log-likelihood. We control the bias of the estimator and rely on optimization theorems to ensure the convergence of a stochastic gradient scheme, easily adapting to a large number of samples.

Keywords. Count data, optimisation, importance sampling, gradient descent, natural exponential family

1 Model

Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip}) \in \mathbb{N}^p$ be some high-dimensional observation vectors of counts, for individual i varying in $1 \leq i \leq n$. We consider a model relying on a multivariate centered scaled Gaussian low dimensional latent variable $\mathbf{W}_i \in \mathbb{R}^q$ ($q \ll p$) linked to the observations $\mathbf{Y}_i$ through a linear function $f_{\theta,i}$ ($1 \leq i \leq n$) with $\theta \in \mathbb{R}^d$ a vector of parameters. Conditionally on $\mathbf{Z}_i = f_{\theta,i}(\mathbf{W}_i) \in \mathbb{R}^p$, the distribution of the observations is assumed to belong to the natural exponential family (NEF). Formally,

$$\begin{aligned} \text{latent space} \quad & \mathbf{W}_i \sim^{\text{iid}} \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q), \quad \mathbf{Z}_i = f_{\theta,i}(\boldsymbol{W}_i) = \mathbf{C}\mathbf{W}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{O}_i, \\ \text{observation space} \quad & p_\theta\left(Y_{ij}|Z_{ij}\right) = \exp(Y_{ij}Z_{ij} - A(Z_{ij}) - h(Y_{ij})), \quad 1 \leq j \leq p\,, \end{aligned} \tag{1}$$

where h and A are real-valued functions with A convex and differentiable and $\theta = (\boldsymbol{C}, \boldsymbol{\beta})$ with $\mathbf{C} \in \mathbb{R}^{p\times q}, \mathbf{X}_i \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^{m\times p}, \mathbf{O}_i \in \mathbb{R}^p$. The parameter $\theta \in \mathbb{R}^d$ with $d = p(q+m)$ is not identifiable as multiplying $\boldsymbol{C}$ by an orthogonal matrix leaves the model unchanged and only $(\boldsymbol{\Sigma}, \boldsymbol{\beta})$ with $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^\top$ is identifiable. We consider 2 distributions inside the NEF, namely the Poisson and Binomial distributions. Every function is known and only the parameter θ is unknown, which is to be estimated from data $(\mathbf{Y}_i)_{1\leq i\leq n}$.

2 Estimation

A natural strategy to estimate θ is by maximizing the log-likelihood function $\ell(\cdot)$ defined as the finite sum of the log-likelihood of each observation:

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(\mathbf{Y}_i). \tag{2}$$

Proposition 2.1. *For all $1 \leq i \leq n$, $\theta \mapsto \ell_i(\theta)$ is $\mathcal{C}^1$.*

2.1 Biased Stochastic Gradient Descent

Given $T \geq 1, \eta > 0$ and $\theta^{(0)} \in \mathbb{R}^d$, one can recursively define $\theta^{(t)}$ via Stochastic Gradient Descent:

$$\theta^{(t+1)} = \theta^{(t)} + \eta\widehat{g}^{(t)} \tag{SGD}$$

where, at each iteration $t \geq 0$, $\widehat{g}^{(t)}$ is a (possibly biased) estimator of $\nabla_\theta \ell(\theta^{(t)})$. An estimator is suggested in the next paragraph.

Constructing the Estimator Let $N \geq 1$ be a number of Monte-Carlo samples, $\{\pi(\cdot, \theta, i)\}_{\theta\in\mathbb{R}^d}$ a family of positive density on $\mathbb{R}^q$ with $1 \leq i \leq n$. An index $i(t) \sim \text{Unif}\{1, \cdots, n\}$ is sampled and conditionnally on $\theta^{(0)}, \cdots, \theta^{(t)}, i(t)$, $\left(\mathbf{V}_k^{(t)}\right)_{1\leq k\leq N} \overset{\text{iid}}{\sim} \pi^{(t)}$ with $\pi^{(t)} \triangleq \pi(\cdot, \theta^{(t)}, i(t))$ and

$$\widehat{g}^{(t)} \triangleq \sum_{k=1}^{N} \omega_k^{(t)} \nabla_\theta \log p_{\theta^{(t)}}(\mathbf{Y}_{i(t)}, \mathbf{V}_k^{(t)}), \text{ where } \omega_k^{(t)} = \frac{\rho_k^{(t)}}{\sum_{\ell=1}^{N}\rho_\ell^{(t)}} \text{ with } \rho_k^{(t)} = \frac{p_{\theta^{(t)}}(\mathbf{Y}_{i(t)}, \mathbf{V}_k^{(t)})}{\pi^{(t)}(\mathbf{V}_k^{(t)})}. \tag{3}$$

We denote $G^{(t)} = \otimes_{k=1}^{N} \pi^{(t)}$ and $\mathbf{V}^{(t)} \triangleq \left(\mathbf{V}_k^{(t)}\right)_{1 \leq k \leq N}$ so that conditionnaly on $\theta^{(0)}, \ldots, \theta^{(t)}$ and $i(t)$, $V^{(t)} \sim G^{(t)}$. The resulting algorithm when Eq. (3) is plugged in SGD is presented in Algorithm 1.

Algorithm 1: Pseudo code SGIS

Input $\theta^{(0)} \in \mathbb{R}^d$ initial point, $T \geq 1$ number of iterations, $\eta > 0$ learning rate, $N \geq 1$ number of Monte-Carlo samples.
Output $\theta^{(0)}, \ldots, \theta^{(T-1)}$
for $t = 0 \ldots T-1$ **do**
  Sample $i(t) \sim \text{Unif}\{1, \ldots, n\}$
  Sample $\mathbf{V}_k \sim \pi^{(t)} (1 \leq k \leq N)$
  Compute $\widehat{g}^{(t)}$ as in Eq. (3)
  Update $\theta^{(t+1)} = \theta^{(t)} + \eta \widehat{g}^{(t)}$
end

3 Convergence guarantees

A differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be $L-$smooth for $L \geq 0$ if for all $\theta, \theta' \in \mathbb{R}^d, \|\nabla_\theta f(\theta) - \nabla_\theta f(\theta')\| \leq L \|\theta - \theta'\|$. The following theorem states sufficient conditions to ensure convergence of SGD.

Theorem 3.1. *[Ajalloeian and Stich [2021]] Let $\epsilon > 0, \theta^{(0)} \in \mathbb{R}^d$ and assume ℓ is $L-$smooth. If there exists $\xi, \sigma > 0$ such that for all $t \geq 1$, $G^{(t)}$ is chosen such that*

$$\mathbb{E}_{G^{(t)}} \left[\left\| \widehat{g}^{(t)} - \nabla_\theta \ell \left(\theta^{(t)} \right) \right\|^2 \right] \leq \sigma$$
$$\left\| \mathbb{E}_{G^{(t)}} \left[\widehat{g}^{(t)} \right] - \nabla_\theta \ell \left(\theta^{(t)} \right) \right\|^2 \leq \xi,$$

then the sequence $\left(\theta^{(t)}\right)_{0 \leq t \leq T-1}$ defined by Algorithm 1 with $\eta = \min\left(\frac{1}{L}, \frac{\epsilon+\zeta}{2L\sigma}\right)$ and $T \geq K\left(\frac{1}{\epsilon+\zeta} + \frac{\sigma}{\epsilon^2+\zeta^2}\right)$ for some $K > 0$ satisfies

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \nabla_\theta \ell \left(\theta^{(t)} \right) \right\|^2 \right] \leq \tilde{K} \left(\epsilon + \zeta \right), \tag{4}$$

for some $\tilde{K} > 0$.

The bias of the suggested estimator 3 can be monitored thanks to the following theorem.

Theorem 3.2. *[Agapiou et al. [2017]] For all $t \geq 0$ and $1 \leq i \leq n$ we have*

$$\mathbb{E}_{G^{(t)}} \left[\left\| \widehat{g}^{(t)} - \nabla_\theta \ell_{i(t)} \left(\theta^{(t)} \right) \right\|^2 \right] \leq \frac{M_{\pi^{(t)}, \theta^{(t)}}}{N} \tag{5}$$

and

$$\left\|\mathbb{E}_{G^{(t)}}\left[\widehat{g}^{(t)}\right] - \nabla_\theta \ell_{i(t)}\left(\theta^{(t)}\right)\right\|^2 \leq \frac{\tilde{M}_{\pi^{(t)},\theta^{(t)}}}{N} \tag{6}$$

with $M_{\pi^{(t)},\theta^{(t)}}$ *and* $\tilde{M}_{\pi^{(t)},\theta^{(t)}}$ *are constant detailed in the appendix.*

In the following, we assume that $\theta \in \mathcal{X}$ with $\mathcal{X}$ a compact convex subset of $\mathbb{R}^d$. Consider π a distribution on $\mathbb{R}^q$ and $\psi : \mathbb{R}^q \mapsto \mathbb{R}^p$, we denote $\psi \in \mathcal{L}^r(\pi)$ if

$$\mathbb{E}_\pi\left[\|\psi(W)\|_1^r\right] < \infty.$$

Corollary 3.3. *[Convergence guarantees]*

Let $1 \leq i \leq n$ *a parametric family of positive density on* $\mathbb{R}^q$ *and* $\lambda_i \in \mathbb{R}_+$ *such that* ${}^{p_{\theta^{(t)}}(\cdot|\mathbf{Y}_i)}/_{\pi(\cdot,\theta^{(t)},i)} \leq \lambda_i$ *for all* $t \geq 0$. *If* $\mathbb{E}_{\pi(\cdot,\theta,i)}\left[\|\nabla_\theta \log\left(p_\theta(\mathbf{Y}_i, \mathbf{W})\right)\|_1^4\right]$ *is finite then assumptions of Theorem 3.1 are verified for Algorithm 1 if the iterates* $(\theta^{(t)})_{t\geq 1}$ *are assumed to belong to the compact convex subset* $\mathcal{X}$. *Moreover, the bias asymptotically vanishes as* $\zeta = \frac{\alpha}{N}$ *for some* $\alpha > 0$.

4 Discussion

Theorem 3.1 can be applied only if Theorem 3.2 is valid, which requires the iterates $(\theta^{(t)})_{t\geq 1}$ remains in the compact X, but cannot be proved without loss of generality. An alternative would be to consider a projection step:

$$\theta^{(t+1)} = \theta^{(t)} - P_{\mathcal{X}}\left(\theta^{(t)} - \eta\widehat{g}^{(t)}\right)$$

where $P_{\mathcal{X}}(\cdot)$ denotes the projection on $\mathcal{X}$, but this does not fall into the setting of Theorem 3.1 so that an adjustment of this Theorem would be preferable and is currently under investigation.

Simulations have been performed and comparison with Chiquet et al. [2018] are promising.

References

Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of sgd with biased gradients, 2021.

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost, 2017.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674 – 2698, 2018.