A probabilistic model for coverage bias estimation and CNV detection

Mehrtash Babadi,* David Benjamin,† and Samuel K. Lee‡
Broad Institute, 75 Ames Street, Cambridge, MA 02142
(Dated: September 9, 2016)

These notes exclusively cover the target coverage model in the GATK CNV pipeline.

I. INTRODUCTION

We wish to address several goals in this section:

- Connect copy ratio (or copy number) and raw read counts in a single sensible probabilistic model without heuristic data transformation.
- Take into account the Poisson nature of coverage depth, thereby giving less weight to low-coverage targets and separating the inherent variance due to Poisson statistics from experimental noise. We want to use the panel of normals to subtract only the latter.
- Choose the number of principal components to use in an automatic and principled manner.
- Use an algorithm that does not waste time calculating all principal components when we only want the few most significant ones.
- Make a universal panel of normals for both sexes by taking into account both autosomal and allosomal targets. This requires the flexibility to handle samples with "missing data".
- Correct for CNV events that occur in the panel of normals.
- Regularize the model property to ensure biological CNV events and laboratory biases are separable from each other, in particular when dealing with a small number of samples.

A. Notation

We use bold symbols for vectors and matrices (e.g. $\mathbf{n}$) and the corresponding regular symbols when the indices are explicitly written (e.g. n_{st}). We use the notation $\mathbf{n}_s$ to refer to the s row vector of the full matrix $\mathbf{n}$. Roman indices are used for sample and target indices whereas Greek indices are reserved for latent space indices.

II. THE MODEL

Suppose we have vectors of read counts over a set of T targets for S samples, $\mathbf{n}_s$, $s = 1 \dots S$ where n_{st} is the coverage of sample s at target t. In order to include both sexes on an equal footing, we further define a "germline ploidy matrix" $\mathcal{P}_{st}$ such that $\mathcal{P}_{st}$ is the germline ploidy[1] of target t of sample s. We imagine that laboratory conditions for a particular sample yielding an underlying bias vector $\mathbf{b}_s$, where $e^{b_{st}}$ is the propensity of target t to be captured, sequenced, and mapped in the preparation of sample s. Suppose also that sample s has an average depth d_s and a vector of copy numbers $\mathbf{c}_s$, where the latent variable c_{st} is the copy number of sample s at target t. Our model for read counts is:

$$n_{st} \sim \text{Poisson}(d_s \mathcal{P}_{st} c_{st} e^{b_{st}}) \tag{1}$$

*Electronic address: mehrtash@broadinstitute.org
†Electronic address: davidben@broadinstitute.org
‡Electronic address: slee@broadinstitute.org
[1] For human autosomal targets, $\mathcal{P}_{st} = 2$ for both sexes. In female samples, $P_{st} = 2$ for X chromosome targets and $P_{st} = 0$ for Y chromosome targets. Finally, $P_{st} = 1$ for X and Y chromosomes in male samples

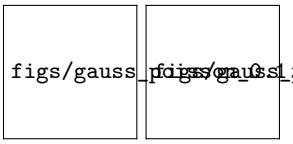


FIG. 1: Gaussian approximation to the Poisson likelihood (see Eq. 3). The left and right panels show $\text{Poisson}(n|\alpha\, e^b)$ and $n^{-1}\mathcal{N}(b|\ln(n/\alpha), n^{-1})$, respectively for $\alpha = 0.1$ (top) and $\alpha = 10.0$ (bottom). The black lines show $b = \ln(n/\alpha)$ the maximum likelihood bias estimate. The Gaussian approximation breaks down at $n = 0$ (no coverage). It also slightly overestimates the variance at small n. Otherwise, it is an excellent approximation.

We can achieve many of the goals listed above by performing probabilistic principal component analysis (PCA) not on directly $\mathbf{n}$, but rather on $\mathbf{b}$. One one hand, the Poisson parameters must be positive and therefore, $\exp(\mathbf{b})$ is a well-defined parametrization of the multiplicative bias. On the other hand, a Gaussian model for $\mathbf{b}$ implies a log-normal distribution for $\exp(\mathbf{b})$ which is indeed the expected distribution when the multiplicative bias arises from several independent sources according to the central limit theorem[2]. We model $\mathbf{b}_s$ as:

$$\begin{aligned} \mathbf{z}_s &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \mathbf{b}_s &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_s + \mathbf{m}, \mathbf{\Psi}), \end{aligned} \tag{2}$$

where $\mathbf{z}_s \in \mathbb{R}^D$ is a low-dimensional latent vector representing laboratory conditions, $\mathbf{W} \in \mathbb{R}^{T\times D}$ is a linear map from latent space to target space, $\mathbf{m} \in \mathbb{R}^T$ is the vector of mean biases, and $\mathbf{\Psi} \in \mathbb{R}^{T\times T}$ is a diagonal matrix of residual variance not explained by the latent features. We approximate the Poisson as a Gaussian and expand the argument of the Gaussian exponential about the mode of b_{st} to quadratic order to obtain:

$$\text{Poisson}(n_{st}|d_s\mathcal{P}_{st}c_{st}e^{b_{st}}) \simeq \Sigma_{st}\,\mathcal{N}(b_{st}|m_{st}, \Sigma_{st}), \tag{3}$$

where:

$$\begin{aligned} m_{st} &\equiv \ln(n_{st}/\mathcal{P}_{st}) - \ln(c_{st}) - \ln(d_s), \\ \Sigma_{st} &\equiv 1/n_{st}. \end{aligned} \tag{4}$$

Note that Σ_{st} can be thought of as the width of the distribution of b_{st} about its maximum likelihood estimate such that in the limit $n_{st}, d_s \to \infty$, $\text{Poisson}(n_{st}|d_s c_{st} e^{b_{st}}) \to \delta(b_{st} - b^*_{st})$ where $b^*_{st} = \lim_{n,d\to\infty} m_{st}$ is the true bias. The above approximation, while being excellent for well-covered targets (see Fig. 1), inevitably breaks down for targets that are uncovered *ex ante* in some samples, such as Y chromosome targets in female samples. To this end, we define a "sample-target mask matrix" M_{st} such that $M_{st} = 0$ if $\mathcal{P}_{st} = 0$, and $M_{st} = 1$ if $\mathcal{P}_{st} \neq 0$, and for each sample-target pair (s, t), we only consider targets where the $M_{st} \neq 0$ in the joint likelihood function. The latter is thus written as:

$$P(n_{st}, b_{st}, z_{st}, d_s|c_{st}, \boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{z}_s|\mathbf{0}, \mathbf{I}) \left[\mathcal{N}(b_{st}|(\mathbf{W}\mathbf{z}_s)_t + m_t, \Psi_t)\,\mathcal{N}(b_{st}|m_{st}, \Sigma_{st})\right]^{M_{st}}, \tag{5}$$

where $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{m}, \mathbf{\Psi})$ denote the parameters we wish to learn. We can integrate out b_{st} readily to obtain[3]:

$$P(n_{st}, z_{st}, d_s|c_{st}, \boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{z}_s|\mathbf{0}, \mathbf{I}) \left[\mathcal{N}(m_{st}|(\mathbf{W}\mathbf{z}_s)_t + m_t, \Psi_{st})\right]^{M_{st}}, \tag{6}$$

where we have defined:

$$\Psi_{st} = \Psi_t + \Sigma_{st}. \tag{7}$$

We model the copy number states (or copy ratios in case of somatic samples) using a finite state hidden Markov model

---

[2] Let $B = \prod_{j=1}^{N_B} B_j$ be the total multiplicative bias where $B_j \in (0, \infty)$ are independent components of the bias. For $N_B \gg 1$, $\ln(B) \sim \mathcal{N}$ and therefore, B has a log-normal distribution.

[3] The integration is easily performed using the identity $\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu_1, \sigma_1^2)\,\mathcal{N}(x|\mu_2, \sigma_2^2)\,\mathrm{d}x = \mathcal{N}(\mu_1|\mu_2, \sigma_1^2 + \sigma_2^2)$.

(HMM). Put together, we obtain:

$$P(\mathbf{n}_s, \mathbf{z}_s, \mathbf{c}_s, d_s|\boldsymbol{\theta}) \propto P_{\text{HMM}}\big(\mathbf{c}_s\big|\boldsymbol{\pi}, \{\mathbf{T}_t\}\big) \prod_t P(n_{st}, z_{st}, d_s|c_{st}, \boldsymbol{\theta}),$$
$$P_{\text{HMM}}\big(\mathbf{c}_s\big|\boldsymbol{\pi}, \{\mathbf{T}_t\}\big) \equiv P(c_{s,0}|\boldsymbol{\pi}) \prod_{t=1}^{T-1} P(c_{s,t}|c_{s,t-1}, \mathbf{T}_t). \tag{8}$$

Here, $\boldsymbol{\pi}$ is a vector denoting the prior probabilities of various copy number states, and $\mathbf{T}_t$ is the transition matrix at target t. We treat HMM parameters as given (hyperparameters).

It is illuminating to study Eq. (6) before we proceed. To this end, we marginalize the bias latent variable $\mathbf{z}_s$ in Eq. (6) to obtain the incomplete-data likelihood function. The final result can be put in a simple form using the Woodbury identity and properties of projection matrices:

$$P(\mathbf{n}|\mathbf{c}, \boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\sum_s (\mathbf{m}_s - \mathbf{m})^T \mathbf{M}_s (\boldsymbol{\Psi} + \boldsymbol{\Sigma}_s + \mathbf{M}_s \mathbf{W}\mathbf{W}^T \mathbf{M}_s)^{-1} \mathbf{M}_s (\mathbf{m}_s - \mathbf{m})\right), \tag{9}$$

where $\mathbf{M}_s$ and $\boldsymbol{\Sigma}_s$ are $T \times T$ diagonal matrices having M_{st} and Σ_{st} in their diagonal entries for $t = 0, \ldots, T-1.$, respectively. The incomplete-data likelihood function poses a Gaussian distribution for $\ln(n_{st})$ (See Eq. 4). The covariance matrix is the diagonal in the target space and is composed of three terms: (1) $\boldsymbol{\Psi}$ denotes target-specific "unexplained" noises, (2) $\mathbf{M}_s\mathbf{W}\mathbf{W}^T\mathbf{M}_s$ denotes the variance explained by the model, and (3) $\Sigma_{st} = 1/n_{st}$ denotes the statistical variance in read counts. This term originates from the Poisson distribution for read counts and can be thought of as a penalty factor that decreases the role of lower-coverage samples in the likelihood.

A. A Laplace regularization scheme for separating latent features from biological CNV events

PCA-like approaches to denoising aim at minimizing the *total variance* of the data by learning and subtracting the contribution of the underlying latent features. In practice, this objective is achieved using either the maximum variance principle (usual PCA) or the maximum likelihood principle on a linear-Gaussian model as explained here. Using either method, when the number of samples largely exceeds the dimension of the latent space, sample-specific variations become immaterial and the true underlying latent features can be learned from the data. However, when the number of samples is comparable to the number of latent features, the statistical power for separating sample-specific variations from mutual variations is significantly reduced.

Let us assume that we have an oracle for the first few major latent features, and that we have already subtracted the contribution arising from these features. Let σ_ℓ^2 be the variance associated with the next leading latent feature. Subtracting this latent feature reduces the total variance by $S\sigma_\ell^2$, where S is the number of samples. Now, if one of the samples has an individual leftover variance of magnitude σ_s^2 such that $\sigma_s^2 \gtrsim S\sigma_\ell^2$, then the maximum variance principle implies choosing the next principal component along the direction of that specific sample. In other words, the procedure erroneously learns a sample-specific signal as a source of noise. Note that this artifact occurs only if $S \lesssim \sigma_s^2/\sigma_\ell^2$.

What is at stake? — There is no obvious theoretical guarantee for the MLE problem for $\boldsymbol{\theta}$ to be convex. In all likelihood, if one of the samples has a large germline CNV event, it may be picked up as a principal component and be interpreted as experimental noise such that the MAP estimator for $\mathbf{c}$ fails to call that CNV event. It is possible that the likelihood function has numerous such local maxima. Therefore, we wish to ensure that sample-specific *nuisances* are not picked up as Gaussian noise, no matter how strong they are. We discuss a number of such approaches in what follows.

(Idea 1) Blind source separation — One remedy is to use a blind source separation approach, such as independent component analysis (ICA), to separate the signal from the noise as a first step, followed by learning the latent features of the noise using PCA. In ICA-like methods, one decomposes the signal into additive subcomponents and minimizes the mutual information between them (or maximizes the non-Gaussianity by taking into account higher moments such as the kurtosis). Even though this method is quite appealing, we follow a more context-specific heuristic approach here.

(Idea 2) Employing a CNV-sensitive regularizer — Fortunately, we have some idea about the spatial structure of the CNV events: they are amplification or attenuation of the read count over several consecutive targets. In the absence of noise, we expect the frequency spectrum of the CNV signal (as obtained by taking a Fourier transform of

m_{st} in t) to be significantly enhanced at spatial frequencies corresponding to the inverse length scale of the size of the CNV event. Similarly, the variation subtracted from sample s, i.e. $\mathbf{W}\mathbf{z}_s$, will exhibit an enhanced spectral power if a CNV event is erroneously picked up. Let $\tilde{f}(k)$ be the Fourier transform of a linear filter that approximately represents a range of CNV events. For example, we may use a midpass filter such as:

$$\tilde{f}(k) = \begin{cases} 1 & k_l \leq k^* \leq k_h, \\ \\ 0 & k^* > k_h \text{ or } k^* < k_l, \end{cases} \tag{10}$$

Here, $k^* = \min(k, T-1-k)$, $k_l \sim \lfloor T/\ell_{\max} \rfloor$ and $k_l \sim \lfloor T/\ell_{\min} \rfloor$, where $\ell_{\min}$ and $\ell_{\max}$ denote roughly the minimum and maximum length of the CNV events in the units of targets. The filtered spectral power of the noise in sample s is given as:

$$\kappa_s \equiv \sum_{k=0}^{T-1} \tilde{f}(k) \left| \mathrm{FFT}\left[\mathbf{W}\mathbf{z}_s\right]_k \right|^2 = \frac{1}{T} \sum_{t,t'=0}^{T-1} F_{tt'}\, W_{t\mu}\, W_{t'\nu}\, z_{s\mu}\, z_{s\nu}, \tag{11}$$

where $F_{tt'} = f(t-t') \equiv T^{-1} \sum_{k=0}^{T-1} e^{2\pi i k(t-t')/T} \tilde{f}(k)$ is the inverse DFT of $\tilde{f}(k)$. Now, in order to avoid picking up event-like variations as noise, we simply penalize variations with large κ. To this end, we regularize the coverage likelihood function Eq. (6) with the following Laplace penalty:

$$R_f \equiv \exp\left(-\frac{\lambda}{2} \sum_{s=1}^{S} \kappa_s \right) = \exp\left(-\frac{\lambda}{T} \sum_{s=1}^{S} \sum_{t,t'=0}^{T-1} f(t-t')\, W_{t\mu}\, W_{t'\nu}\, z_{s\mu}\, z_{s\nu} \right). \tag{12}$$

We define the matrix $F_{t,t'} \equiv f(t-t')/T$ for notational convenience. The regularizer "kicks in" when $\lambda \sim \Psi^{-1}$, as it can be inferred directly from Eq. (27). One may initially choose $\lambda \sim 1000\, \Psi^{-1}$ and progressively relax it as the optimal solution is approached.

B. Automatic relevance determination (ARD) prior on W

The true dimension of the latent space is not known *a priori*. When abundant data is available (i.e. $S \gg D_{\text{true}}$), this problem can be addressed using the automatic relevance determination (ARD) technique. To this end, one starts with a liberal estimate for D and imposes a Gaussian prior on the rows of $\mathbf{W}$ (principal components):

$$P(\mathbf{W}) \propto \prod_{\mu} \alpha_\mu^{T/2} \exp\left(-\frac{1}{2} \alpha_\mu \sum_t W_{t\mu}^2 \right). \tag{13}$$

If $\alpha_\mu \to \infty$, the latent feature μ is effectively turned off whereas if $\alpha_\mu \to 0$, the flat prior is recovered. The recipe for ARD is to set $\alpha_\mu \simeq 0$ initially while calculating and maximizing the model evidence $P(\mathbf{n}|\{\alpha_\mu\}$ with respect to $\{\alpha_\mu\}$ during the EM algorithm. If $D > D_{\text{true}}$, we expect $D - D_{\text{true}}$ elements of $\{\alpha_\mu\}$ to run to infinity. If $D < D_{\text{true}}$, all of $\{\alpha_\mu\}$ will remain of the same order, signaling the necessity of increasing D.

III. MEAN-FIELD VARIATIONAL EM ALGORITHM

We start by writing the complete-data log likelihood (see Eq. 14a), including the ARD prior and the Laplace regularizer:

$$\begin{aligned} \ln P(\mathbf{n}, \mathbf{c}, \mathbf{z}, \mathbf{d}|\boldsymbol{\theta}, \boldsymbol{\alpha}) = {} & \sum_s \ln P_{\text{HMM}}\big(\mathbf{c}_s \big| \boldsymbol{\pi}, \{\mathbf{T}_t\}\big) - \frac{1}{2} \sum_{st} M_{st} \left\{ \ln \Psi_{st} + \Psi_{st}^{-1} \Big[(\mathbf{W}\mathbf{z}_s)_t^2 + \Delta_{st}^2 - 2\Delta_{st} (\mathbf{W}\mathbf{z}_s)_t \Big] \right\} \\ & - \frac{\lambda}{2} \sum_s \sum_{t,t'} F_{tt'}\, W_{t\mu}\, W_{t'\nu}\, z_{s\mu}\, z_{s\nu} - \frac{1}{2} \sum_s \mathbf{z}_s^T \mathbf{z}_s + \sum_{\mu=1}^{D} \left(\frac{T}{2} \ln \alpha_\mu - \frac{\alpha_\mu}{2} \sum_{t=0}^{T-1} W_{t\mu}^2 \right) + \text{const.}, \end{aligned} \tag{14a}$$

$$\Delta_{st} \equiv \ln(n_{st}/\mathcal{P}_{st}) - \ln c_{st} - \ln d_s - m_t. \tag{14b}$$

We notice that the latent variables $(\mathbf{c}, \mathbf{d})$ are coupled to $\mathbf{z}$ due to the cross term $\Delta\mathbf{W}\mathbf{z}$, and $\mathbf{c}$ is coupled to $\mathbf{d}$ due to Δ^2. These couplings result in an intractable E step. We expect the posterior distribution for $\mathbf{z}$ and $\mathbf{d}$ to be quite sharp near the optimal solution, provided $D_{\text{true}} \ll T$. Therefore, we do not foresee these couplings to play a substantial role, allowing us to utilize an approximate factorized variational ansatz for the latent posterior distribution:

$$\Phi(\mathbf{c}, \mathbf{z}, \mathbf{d}) \equiv P(\mathbf{c}, \mathbf{z}, \mathbf{d}|\mathbf{n}, \boldsymbol{\theta}_{\text{old}}) \simeq \Phi_c(\mathbf{c})\, \Phi(\mathbf{z})\, \Phi(\mathbf{d}). \tag{15}$$

The E step requires solving the following set of mean-field self-consistency equations:

$$\ln \Phi_c(\mathbf{c}) = \mathbb{E}_{\mathbf{z},\mathbf{d}}\big[\ln P(\mathbf{n}, \mathbf{c}, \mathbf{z}, \mathbf{d}|\boldsymbol{\theta})\big], \tag{16a}$$
$$\ln \Phi_z(\mathbf{z}) = \mathbb{E}_{\mathbf{c},\mathbf{d}}\big[\ln P(\mathbf{n}, \mathbf{c}, \mathbf{z}, \mathbf{d}|\boldsymbol{\theta})\big], \tag{16b}$$
$$\ln \Phi_d(\mathbf{d}) = \mathbb{E}_{\mathbf{c},\mathbf{z}}\big[\ln P(\mathbf{n}, \mathbf{c}, \mathbf{z}, \mathbf{d}|\boldsymbol{\theta})\big], \tag{16c}$$

where the posterior expectation values are calculated with respect to the factorized distribution. Writing out the right hand side explicitly, we find:

$$\Phi_c(\mathbf{c}_s) \propto P_{\text{HMM}}\big(\mathbf{c}_s\big|\boldsymbol{\pi}, \{\mathbf{T}_t\}\big) \prod_{t=0}^{T-1} \exp\left[-\frac{1}{2} M_{st} \Psi_{st}^{-1} \big[\ln(n_{st}/\mathcal{P}_{st}) - \ln c_{st} - \mathbb{E}[\ln d_s] - m_t - (\mathbf{W}\mathbb{E}[\mathbf{z}_s])_t\big]^2\right], \tag{17}$$

$$\Phi_z(\mathbf{z}_s) \propto \exp\left[-\frac{1}{2}\, \mathbf{z}_s^T \Big(\mathbf{I} + \mathbf{W}^T \mathbf{M}_s \boldsymbol{\Psi}_s^{-1} \mathbf{W} + \lambda \mathbf{W}^T \mathbf{F} \mathbf{W}\Big)\, \mathbf{z}_s + \mathbf{W}^T \mathbf{M}_s\, \mathbb{E}[\boldsymbol{\Delta}_s]\, \mathbf{z}_s\right], \tag{18}$$

$$\Phi_d(d_s) \propto \exp\left[-\frac{1}{2} \sum_{t=0}^{T-1} M_{st} \Psi_{st}^{-1} \big[\ln(n_{st}/\mathcal{P}_{st}) - \mathbb{E}[\ln c_{st}] - \ln d_s - m_t - (\mathbf{W}\mathbb{E}[\mathbf{z}_s])_t\big]^2\right]. \tag{19}$$

We have completed the squares in the first and the last equation in exchange for different normalization factors (which we do not need to calculate). We observe that: (1) $\Phi_z(\mathbf{z}_s)$ is conveniently Gaussian in $\mathbf{z}_s$; (2) $\Phi_d(d_s)$ admits a Gaussian distribution over $\rho \equiv \ln d$:

$$\Phi_\rho(\rho_s) = e^{\rho_s}\, \Phi_d(e^{\rho_s}) \propto \exp\left[\rho_s - \frac{1}{2} \sum_{t=0}^{T-1} M_{st} \Psi_{st}^{-1} \big[\ln(n_{st}/\mathcal{P}_{st}) - \mathbb{E}[\ln c_{st}] - \rho_s - m_t - (\mathbf{W}\mathbb{E}[\mathbf{z}_s])_t\big]^2\right], \tag{20}$$

which is normalizable and does not require regularization; (3) the factorized distribution for $\mathbf{c}_s$ yields an emission model that is local in the target space. These observations allow us to arrive at a simple recipe for the E step, which can be summarized as the following set of self-consistent mean-field equations:

$$\mathbb{E}\big[\ln d_s\big] = \frac{1 + \sum_t M_{st} \Psi_{st}^{-1} \left[\ln(n_{st}/\mathcal{P}_{st}) - \mathbb{E}[\ln c_{st}] - m_t - (\mathbf{W}\mathbb{E}[\mathbf{z}_s])_t\right]}{\sum_t M_{st} \Psi_{st}^{-1}},$$
$$\mathbb{E}\big[\Delta_{st}\big] = \ln(n_{st}/\mathcal{P}_{st}) - \mathbb{E}\big[\ln c_{st}\big] - \mathbb{E}\big[\ln d_s\big] - m_t, \tag{21a}$$
$$\mathbb{E}\big[\mathbf{z}_s\big] = \mathbf{G}_s \mathbf{W}^T \mathbf{M}_s \boldsymbol{\Psi}_s^{-1} \mathbb{E}\big[\boldsymbol{\Delta}_s\big], \qquad \mathbf{G}_s \equiv \big(\mathbf{I} + \mathbf{W}^T \mathbf{M}_s \boldsymbol{\Psi}_s^{-1} \mathbf{W} + \lambda \mathbf{W}^T \mathbf{F} \mathbf{W}\big)^{-1} \tag{21b}$$
$$\mathbb{E}\big[\ln c_{st}\big] = \sum_c \gamma_{st}(c) \ln(c), \tag{21c}$$

where in the last equation, the summation is over different copy number (or ratio) states, and $\gamma_{st}(c) \equiv P(c_{st} = c|\mathbf{n})$ which can be efficiently found using forward-backward algorithm. In practice, we can set $\mathbb{E}[\mathbf{z}_s] = \mathbb{E}[\ln c_{st}] = 0$ initially and cycle through the mean-field equations for d, z and c in succession until convergence is achieved to a desired degree. Once the self-consistent solution is found, the additional quantities required for calculating the posterior expectation of $\ln P$ are easily found as:

$$\mathbb{E}\big[\mathbf{z}_s \mathbf{z}_s^T\big] = \mathbf{G}_s + \mathbb{E}\big[\mathbf{z}_s\big] \mathbb{E}\big[\mathbf{z}_s\big]^T,$$
$$\mathbb{E}\big[\Delta_{st}^2\big] = \mathbb{E}[\Delta_{st}]^2 + \text{var}\big[\ln c_{st}\big] + \text{var}\big[\ln d_s\big],$$
$$\text{var}\big[\ln c_{st}\big] = \sum_c \gamma_{st}(c) \ln^2(c) - \mathbb{E}\big[\ln c_{st}\big]^2,$$
$$\text{var}\big[\ln d_s\big] = \left(\sum_t M_{st} \Psi_{st}^{-1}\right)^{-1} \tag{22}$$

In the M step, we calculate the expectation value of the complete-data log likelihood with respect to the posterior distribution, and maximize it with respect to $\boldsymbol{\theta}$. The quantity of interest is:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}},\boldsymbol{\alpha}) &= \mathbb{E}_{q(\mathbf{c},\mathbf{z},\mathbf{d})}\big[\ln P(\mathbf{n},\mathbf{c},\mathbf{z},\mathbf{d}|\boldsymbol{\theta})\big] \\ &= -\frac{1}{2}\sum_{st} M_{st}\left\{\ln\Psi_{st} + \Psi_{st}^{-1}\Big[\big(\mathbf{W}\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^T]\mathbf{W}\big)_{tt} + \mathbb{E}\big[\Delta_{st}^2\big] - 2\,\mathbb{E}\big[\boldsymbol{\Delta}_{st}^T\big]\,\mathbf{W}\,\mathbb{E}\big[\mathbf{z}_s\big]\Big]\right\} \\ &\quad -\frac{\lambda}{2}\sum_s \mathrm{Tr}\Big(\mathbf{W}^T\mathbf{F}\mathbf{W}\,\mathbb{E}\big[\mathbf{z}_s\mathbf{z}_s^T\big]\Big) + \sum_{\mu=1}^{D}\left(\frac{T}{2}\,\ln\alpha_\mu - \frac{\alpha_\mu}{2}\sum_{t=0}^{T-1} W_{t\mu}^2\right) + \text{const.}\end{aligned} \tag{23}$$

Maximizing $\mathcal{Q}$ with respect to $\mathbf{m}$ gives:

$$m_t = \left(\sum_s \mathbf{M}_s\boldsymbol{\Psi}_s^{-1}\right)^{-1}\sum_s\left[\mathbf{M}_s\boldsymbol{\Psi}_s^{-1}(\mathbf{m}_s - \mathbf{W}\mathbb{E}[\mathbf{z}_s])\right]. \tag{24}$$

Maximizing $\mathcal{Q}$ with respect to Ψ_t gives:

$$\sum_s M_{st}\left[\frac{1}{\Psi_t + \Sigma_{st}} - \frac{B_{st}}{(\Psi_t+\Sigma_{st})^2}\right] = 0, \tag{25}$$

where:

$$B_{st} = \big(\mathbf{W}\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^T]\mathbf{W}\big)_{tt} + \mathbb{E}\big[\Delta_{st}^2\big] - 2\,\mathbb{E}\big[\boldsymbol{\Delta}_{st}^T\big]\,\mathbf{W}\,\mathbb{E}\big[\mathbf{z}_s\big] \tag{26}$$

The above nonlinear equation must be solved for each target. In practice, we found that the best approach was to use Brent solver for each target. On average, $10 \sim 15$ function calls yields the solution within a 10^{-6} tolerance. Newton's method required approximately 20 evaluations to converge within the same tolerance and was not stable all the time.

Maximizing $\mathcal{Q}$ with respect to $\mathbf{W}$ gives:

$$\sum_\nu \left(Q_{t\mu\nu} + A_{\mu\nu}\right) W_{t\nu} + \lambda\sum_{\nu,t'} Z_{\mu\nu}F_{tt'}W_{t'\nu} = v_{t\mu}, \tag{27}$$

where:

$$\begin{aligned} Q_{t\mu\nu} &= \sum_s M_{st}\Psi_{st}^{-1}\,\mathbb{E}\big[z_{s\mu}z_{s\nu}\big], \\ v_{t\mu} &= \sum_s M_{st}\Psi_{st}^{-1}\mathbb{E}\big[\Delta_{st}\big]\,\mathbb{E}\big[z_{s\mu}\big], \\ A_{\mu\nu} &= \delta_{\mu\nu}\,\alpha_\mu, \\ Z_{\mu\nu} &= \sum_{s=1}^{S}\mathbb{E}[z_{s\mu}z_{s\nu}]. \end{aligned} \tag{28}$$

The sparse structure of the above linear equation allows us to solve it efficiently using Krylov subspace methods (to be discussed later).

Finally, to determine α_μ, we re-exponentiate $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}},\boldsymbol{\alpha})$ and integrate out $\mathbf{W}$ to obtain the evidence for $\mathbf{A}$:

$$P(\mathbf{n}|\mathbf{A}) \propto \prod_\mu \alpha_\mu^{T/2}\int q(\mathbf{W})\,\mathrm{d}\mathbf{W}, \tag{29}$$

where:

$$q(\mathbf{W}) \equiv \exp\left(-\frac{1}{2}\sum_{\mu\nu}\sum_{tt'} W_{t\mu}\left[\left(A_{\mu\nu} + Q_{t\mu\nu}\right)\delta_{tt'} + \lambda F_{tt'}\,Z_{\mu\nu}\right]W_{t'\nu} - \sum_\mu W_{t\mu}v_{t\mu}\right). \tag{30}$$

The ARD coefficients α are determined by maximizing the log evidence, i.e. $\partial/\partial\alpha_\mu \ln P(\mathbf{n}|\mathbf{A}) = 0$ which easily gives:

$$\frac{\alpha_\mu}{T} = \left(\frac{\int \sum_t W_{t\mu}^2 \, q(\mathbf{W}) \, \mathrm{d}\mathbf{W}}{\int q(\mathbf{W}) \, \mathrm{d}\mathbf{W}} \right)^{-1} = \left(\sum_t \left(W_{t\mu}^*\right)^2 + \sum_t \left(\mathbf{A} + \mathbf{Q} + \lambda \mathbf{F} \otimes \mathbf{Z}\right)^{-1}_{\mu t, \mu t} \right)^{-1}, \tag{31}$$

where $\mathbf{W}^*$ is the latest M step value for $\mathbf{W}$. Unfortunately, the presence of the regularizer and the coupling it introduces between the targets makes calculating the required inverse matrix element a numerically challenging problem. Without the regularizer term, the inversion can be performed per-target and only within the low dimensional latent space. One approach is to extract the required matrix elements using stochastic methods. For example, one can efficiently calculate $\mathbf{w}^T \left(\ldots\right)^{-1} \mathbf{v}$ using Krylov subspace methods. If one calculates this scalar quantity for an ensemble of pairs $(\mathbf{v}, \mathbf{w})$ such that $\mathbb{E}[v_{\alpha t} w_{\beta t'}] = \delta_{\alpha\beta}\delta_{\alpha\mu}\delta_{tt'}$, the ensemble average of $\mathbf{w}^T \left(\ldots\right)^{-1} \mathbf{v}$ yields the desired matrix element. Another approach is to disable the regularizer after a number of EM iterations, and enable the ARD prior instead.

A. On the emission probability

The emission probability for the HMM is local in target and sample indices, and can be read from Eq. (17):

$$P_{\mathrm{em}}(n_{st}|c_{st}) = \mathcal{N}_{st} \exp\left[-\frac{1}{2} \Lambda_{st} \left(\ln n_{st} - \mu_{st} \right)^2 \right], \tag{32}$$

where we have defined:

$$\begin{aligned} \Lambda_{st} &= M_{st} \Psi_{st}^{-1}, \\ \mu_{st} &= \ln(c_{st}) + \ln(\mathcal{P}_{st}) + \mathbb{E}[\ln d_s] + m_t + (\mathbf{W}\mathbb{E}[\mathbf{z}_s])_t, \end{aligned} \tag{33}$$

and $\mathcal{N}_{st}$ is a normalization factor. We will drop the st indices for brevity hereafter in this section. The emission function must be a proper probability function, i.e. it must sum to unity by considering all possible read counts. Consequently, the (inverse) normalization factor is given by:

$$\mathcal{N}^{-1} = \sum_{n=1}^{\infty} \exp\left[-\frac{\Lambda}{2} \left(\ln n - \mu \right)^2 \right]. \tag{34}$$

We have left out $n = 0$ from the summation since we implicitly assume such targets are masked out[4]. An excellent approximation in cases where target coverage is high is to replace the discrete sum with an integral:

$$\begin{aligned} \mathcal{N}^{-1} &\simeq \int_1^\infty \mathrm{d}n \, \exp\left[-\frac{\Lambda}{2} \left(\ln n - \mu \right)^2 \right] \\ &= \int_0^\infty \mathrm{d}x \, \exp\left[-\frac{\Lambda}{2} \left(x - \mu \right)^2 + x \right] = \exp\left(\mu + \frac{1}{2\Lambda} \right) \int_0^\infty \mathrm{d}x \, \exp\left[-\frac{\Lambda}{2} \left(x - \mu - \Lambda^{-1} \right)^2 \right] \\ &= \sqrt{\frac{2\Lambda}{\pi}} \, \exp\left(\mu + \frac{1}{2\Lambda} \right) \mathrm{erfc}\left(-\frac{\Lambda\mu + 1}{\sqrt{2\Lambda}} \right). \end{aligned} \tag{35}$$

In summary, the log emission probability is given by:

$$\ln P_{\mathrm{em}}(n|c) \simeq \frac{1}{2} \ln\left(\frac{\pi}{2\Lambda} \right) - \mu - \frac{1}{2\Lambda} - \ln \mathrm{erfc}\left(-\frac{\Lambda\mu + 1}{\sqrt{2\Lambda}} \right) - \frac{1}{2} \Lambda \left(\ln n - \mu \right)^2. \tag{36}$$

Note that c implicitly appears in μ. Also, we recall that $\mathrm{erfc}(-x) = 1 + \mathrm{erf}(x)$ which is useful in numerics.

---

[4] Only unmasked targets must be passed to the HMM.

B. Efficient computational of expressions involving the regularizer

Calculating $\mathbf{W}^T\mathbf{F}\mathbf{W}$ — Since $\mathbf{F}$ is not diagonal in the target space, a naive matrix multiplication implies a multiplication complexity of $\mathcal{O}(D^2T^2)$ for the new term. However, this complexity can be reduced to a manageable $\mathcal{O}(D^2T\log T)$ using FFT:

$$(\mathbf{W}^T\mathbf{F}\mathbf{W})_{\mu\nu} = \sum_{t=0}^{T-1} W_{\mu t}\,\mathrm{FFT}_t^{-1}\left[\sum_{k=0}^{T-1}\tilde{f}(k)\,\mathrm{FFT}_k[W_{t\nu}]\right]. \tag{37}$$

Solving the M step equation for W — Fortunately, the linear operator $\mathbf{Q}+\mathbf{A}+\lambda\mathbf{Z}\otimes\mathbf{F}$, is the sum of two sparse operators: $\mathbf{Q}+\mathbf{A}$ is diagonal in the target space, and $\mathbf{Z}\otimes\mathbf{F}$ is diagonal in Fourier space ($\mathbf{Z}$ acts on the latent space, $\mathbf{F}$ acts on the target space). Both $\mathbf{Q}+\mathbf{A}$ and $\mathbf{Z}\otimes\mathbf{F}$ are dense in the latent space, but this space has a low dimensionality and is not prohibitive in numerics. Eq. (27) can be solved very efficiently using preconditioned iterative Krylov space solvers such as conjugate gradients (CG) or generalized minimal residual (GMRES.) A decent preconditioner can be constructed by taking a target average of $Q_{t\mu\nu}$:

$$\mathbf{\Lambda} \equiv \tilde{\mathbf{Q}} + \mathbf{A} + \lambda\mathbf{Z}\otimes\mathbf{F}, \qquad \tilde{\mathbf{Q}} = \frac{1}{T}\sum_t \mathbf{Q}_t. \tag{38}$$

Note that $\mathbf{\Lambda}$ is now easily invertible in the Fourier space. In iterative methods, we only need to be able to calculate $\mathbf{\Lambda}^{-1}\mathbf{W}$ for arbitrary $\mathbf{W}$. The complexity for this is $\mathcal{O}(D^3T\log T)$ using FFT:

$$\left(\mathbf{\Lambda}^{-1}\mathbf{W}\right)_{t\mu} = \mathrm{FFT}_t^{-1}\left[\left(\tilde{\mathbf{Q}}+\mathbf{A}+\lambda\tilde{f}(k)\mathbf{Z}\right)^{-1}\mathrm{FFT}_k[\mathbf{W}_t]\right]. \tag{39}$$

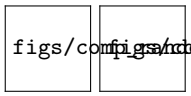
Note that if target-to-target variance $\mathbf{Q}_t$ is small (which is the case if the targets have a comparable degree of unexplained variance), $\mathbf{\Lambda}^{-1}\mathbf{v}$ is an excellent approximate solution to Eq. (27) and can be used as a starting point. In practice, we found preconditioned CG iterations to converge within an error tolerance of 10^{-6} within less than 10 steps. The complexity of each CG step is also $\mathcal{O}(D^3T\log T)$.

IV. RESULTS

In this section, we present the result of the algorithm on synthetic coverage data where the ground truth is known (this section must be eventually supplemented with real data). We synthesize the data according to Eq. (1) along with random duplication events of varying lengths. We choose $T = 4000$ targets, $D = 10$ true latent variables, $S = 100$ samples, mean read depth d uniformly sampled from $[50, 1000]$, mean bias $m_t \sim \mathcal{N}(0, 1)$, eigenvalues of the covariance matrix $\mathbf{W}\mathbf{W}^T$ uniformly sampled from $[0, 10]$, and residual variance Ψ_t uniformly sampled from $[0.01, 0.05]$. Finally, the length of CNV events are randomly sampled from $[50, 500]$ targets.

Fig. 2 compares PCA denoising against our probabilistic model with different features turned on/off (ARD, CNV event regularization) for random and correlated events, respectively. It is clearly observed that the regularized model retains all of the events even when the number of latent features chosen is greater than the true number.

$D = 10$



$D = 20$

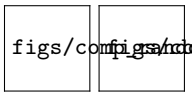


FIG. 2: Comparison of PCA with the probabilistic coverage model in different modes. Top two rows: $D = 10$; random events, correlated events. Bottom two rows: $D = 20$; random events, correlated events.