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: August 8, 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

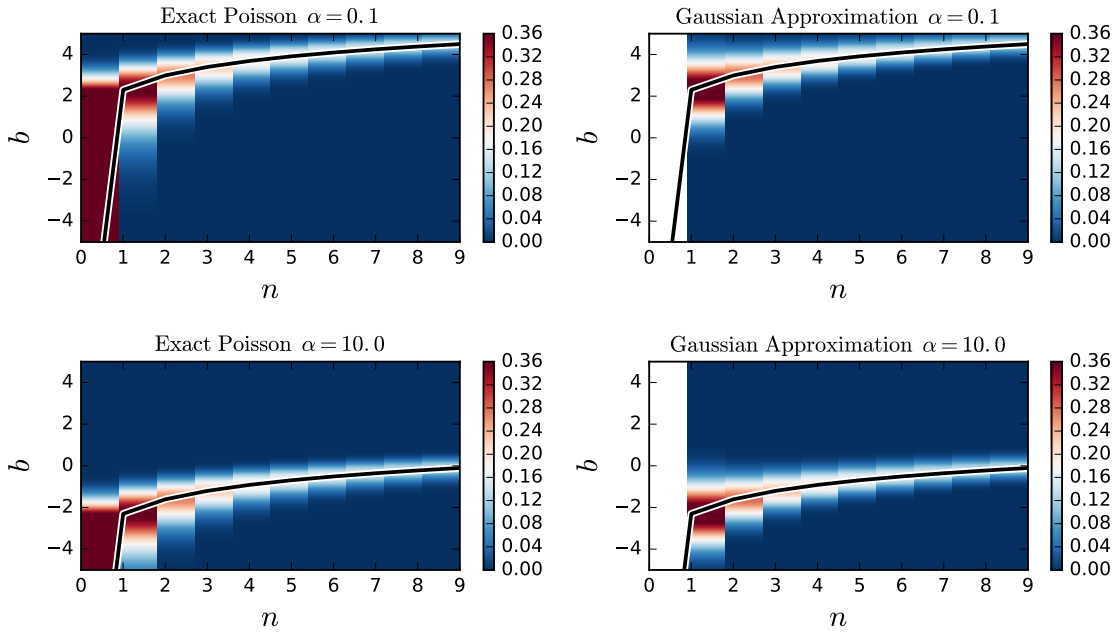

FIG. 1: Gaussian approximation to the Poisson likelihood (see Eq. 3). The left and right panels show $\text{Poisson}(n|\alpha\, e^b)$ and $n^{-1}\mathcal{N}(b|\ln(n/\alpha), n^{-1})$, respectively for $\alpha = 0.1$ (top) and $\alpha = 10.0$ (bottom). The black lines show $b = \ln(n/\alpha)$ the maximum 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

[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.

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 $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{m}, \boldsymbol{\Psi})$ 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 (HMM). Put together, we obtain:

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

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 Σ_s are $T \times T$ diagonal matrices having M_{st} and Σ_{st} in their diagonal entries, 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) $\boldsymbol{\Psi}$ target-specific "unexplained" noises, (2) $\mathbf{M}_s\mathbf{W}\mathbf{W}^T\mathbf{M}_s$ denotes the "explained" variance due to laboratory conditions, and (3) $\Sigma_{st} = 1/n_{st}$ denotes the statistical variance in read counts, it has roots in 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

---

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

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:

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

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 will discuss a proper choice of λ later. We also define the matrix $F_{t,t'} \equiv f(t-t')/T$ for later sections.

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 $\mathbf{W}$:

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

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 initially set $\alpha_\mu \simeq 0$ while calculating and maximizing the model evidence $P(\mathbf{n}|\{\alpha_\mu\}$ with respect to $\{\alpha_\mu\}$. If $D > D_{\text{true}}$, we expect $D - D_{\text{true}}$ elements of $\{\alpha_\mu\}$ to run to infinity. If D_{true}, all of $\{\alpha_\mu\}$ will remain of the same order, signaling the necessity of increasing D.

III. MEAN-FIELD VARIATIONAL EM ALGORITHM

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

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

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

We notice that the latent variables $(\mathbf{c},\mathbf{d})$ are coupled to $\mathbf{z}$ due to the cross term $\Delta\mathbf{W}\mathbf{z}$, resulting in an intractable E step. Intuitively, we expect the posterior distribution of $\mathbf{z}$ to be quite narrow near the optimal solution provided $D_{\rm true} \ll T$. Therefore, we do not foresee the coupling to play a substantial role. This motivates the introduction of a factorized variational ansatz for the latent posterior distribution:

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

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:

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

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

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

Note that 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; (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 $\mathbf{c}_s$ implies an emission model that is local in the target space. These observations allow us to arrive at a simple recipe for the E step. We summarize the mean-field equations as follows:

$$\mathbb{E}\big[\ln d_s\big] = \frac{1 + \sum_t M_{st}\Psi_{st}^{-1}\left[\ln(n_{st}/\mathcal{P}_{st}) - \mathbb{E}[\ln c_{st}] - m_t - (\mathbf{W}\mathbb{E}[\mathbf{z}_s])_t\right]}{\sum_t M_{st}\Psi_{st}^{-1}},$$

$$\mathbb{E}\big[\Delta_{st}\big] = \ln(n_{st}/\mathcal{P}_{st}) - \mathbb{E}\big[\ln c_{st}\big] - \mathbb{E}\big[\ln d_s\big] - m_t, \tag{21a}$$

$$\mathbb{E}\big[\mathbf{z}_s\big] = \mathbf{G}_s\mathbf{W}^T\mathbf{M}_s\boldsymbol{\Psi}_s^{-1}\mathbb{E}\big[\boldsymbol{\Delta}_s\big], \qquad \mathbf{G}_s \equiv \big(\mathbf{I} + \mathbf{W}^T\mathbf{M}_s\boldsymbol{\Psi}_s^{-1}\mathbf{W} + \lambda\mathbf{W}^T\mathbf{F}\mathbf{W}\big)^{-1} \tag{21b}$$

$$\mathbb{E}\big[\ln c_{st}\big] = \sum_c \gamma_{st}(c)\ln(c), \tag{21c}$$

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 message passing method. In practice, we can set $\mathbb{E}[\mathbf{z}_s] =$

$\mathbb{E}[\ln c_{st}] = 0$ and cycle through the mean-field equations for d, z and c in succession until convergence is achieved. Once the self-consistent is found, the additional expectation values required for calculating the posterior expectation of $\ln P$ are easily found as:

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

In the M step, we calculate the expectation value of the complete-data log likelihood with respect to the posterior estimate of $\mathbf{z}_s$. Save for terms independent of the model parameters, the result is:

$$\mathcal{L} = -\frac{1}{2}\sum_{st}\left\{M_{st}\ln\Psi_{st} + M_{st}\Psi_{st}^{-1}\left[\left(\mathbf{W}\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^T]\mathbf{W}^T\right)_{tt} + 2(m_t - m_{st})\left(\mathbf{W}\mathbb{E}[\mathbf{z}_s]\right)_t + (m_t - m_{st})^2\right]\right\}, \tag{23}$$

The stationarity condition for $\mathcal{L}$ with respect to $\mathbf{m}$ gives:

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

The stationarity condition 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:

$$B_{st} = \left(\mathbf{W}\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^T]\mathbf{W}^T\right)_{tt} + 2(m_t - m_{st})\left(\mathbf{W}\mathbb{E}[\mathbf{z}_s]\right)_t + (m_t - m_{st})^2 \tag{26}$$

The above nonlinear equation must be solved for each target, which is a computationally demanding task for a large number of targets. If this is to be avoided, we offer two approximation schemes:

(Scheme 1) Assuming small sample to sample variations in Σ_{st}: Had Σ_{st} been constant, then Eq. (25) would have the following simple solution:

$$\Psi_t^{\mathrm{approx}} = \langle\!\langle\mathbf{B}\rangle\!\rangle_t - \bar{\Sigma}_t, \tag{27}$$

where $\bar{\Sigma}_t$ is the sample-independent value of Σ_{st}, and we have defined the double angle bracket average as:

$$\langle\!\langle\mathbf{B}\rangle\!\rangle_t \equiv \frac{\sum_s M_{st}B_{st}}{\sum_s M_{st}}. \tag{28}$$

It is tempting to replace $\bar{\Sigma}_t$ with its M-averaged value. However, a more principled approach is to choose $\bar{\Sigma}_t$ such that the approximation solution given in (28) is as close to the exact solution as possible. To this end, we assume $\Sigma_{st} = \bar{\Sigma}_t + (\Sigma_{st} - \bar{\Sigma}_t)$ such that $|\Sigma_{st} - \bar{\Sigma}_t| \ll \bar{\Sigma}_t$, expand Eq. (25) in $\Sigma_{st} - \bar{\Sigma}_t$ to linear order, and require that Ψ_t^{approx} represents the exact solution. This procedure yields:

$$\bar{\Sigma}_t = \frac{2\langle\!\langle\mathbf{B}\mathbf{\Sigma}\rangle\!\rangle_t - \langle\!\langle\mathbf{\Sigma}\rangle\!\rangle_t\langle\!\langle\mathbf{B}\rangle\!\rangle_t}{\langle\!\langle\mathbf{B}\rangle\!\rangle_t}. \tag{29}$$

Plugging this result back in Eq. (28), we find:

$$\Psi_t^{\mathrm{approx}} = \langle\!\langle\mathbf{B}\rangle\!\rangle_t + \langle\!\langle\mathbf{\Sigma}\rangle\!\rangle_t - 2\,\frac{\langle\!\langle\mathbf{B}\mathbf{\Sigma}\rangle\!\rangle_t}{\langle\!\langle\mathbf{B}\rangle\!\rangle_t}. \tag{30}$$

This solution is accurate to linear order in deviations of Σ_{st} about $\bar{\Sigma}_t$.

(Scheme 2) Newton iterations: The complexity of solving Eq. (25) numerically is not too high given that the Hessian matrix is diagonal. Expanding $\mathcal{L}$ about $\mathbf{\Psi}_0$, we find:

$$\mathcal{L}(\mathbf{\Psi}) = \mathcal{L}(\mathbf{\Psi}_0) + \alpha_t \left(\Psi_t - \Psi_{0,t}\right) + \frac{1}{2}\,\beta_t \left(\Psi_t - \Psi_{t,0}\right)^2 + \ldots \tag{31}$$

where:

$$\begin{aligned}
\alpha_t &= \frac{\partial \mathcal{L}}{\partial \Psi_t} = -\frac{1}{2}\sum_s M_{st}\left[\frac{1}{\Psi_t + \Sigma_{st}} - \frac{B_{st}}{(\Psi_t + \Sigma_{st})^2}\right], \\
\beta_t &= \frac{\partial^2 \mathcal{L}}{\partial \Psi_t^2} = +\frac{1}{2}\sum_s M_{st}\left[\frac{1}{(\Psi_t + \Sigma_{st})^2} - \frac{2B_{st}}{(\Psi_t + \Sigma_{st})^3}\right].
\end{aligned} \tag{32}$$

The Newton's approximate solution is therefore:

$$\Psi_{t,1} = \Psi_{t,0} - \frac{\alpha_t(\mathbf{\Psi}_0)}{\beta_t(\mathbf{\Psi}_0)}. \tag{33}$$

One may start iterations using the result of Scheme 1 as the initial guess and continue until convergence.

Remark: In practice, when sample-to-sample variance of Σ was large (e.g. read depths varying from 50 to 1000 randomly), we noticed that the best approach was to use Brent root finding for each target. On average, 10 function calls yields the solution within a 10^{-6} tolerance. Newton's method required approximately 20 evaluations of α_t and β_t to converge within the same tolerance. Also, we found that the most robust scheme was to start from $\Psi_t = 0$ rather than using Eq. (30).

In the M step equation for $\mathbf{W}$, we may incorporate an automatic relevance determination (ARD) prior:

$$P(\mathbf{W}) = \prod_\mu \left(\frac{\alpha_\mu}{2\pi}\right)^{T/2} \exp\left(-\frac{1}{2}\alpha_\mu \sum_t W_{t\mu}^2\right). \tag{34}$$

If $\alpha_\mu \to \infty$, the latent feature μ is effectively turned off. Thus we can initially choose a liberal estimate of D and the model will automatically become more parsimonious. The M step log likelihood times the ARD prior depend on the t-th row of $\mathbf{W}$ as:

$$-\frac{1}{2}\left(-\sum_\mu \ln \alpha_\mu + \sum_{\mu\nu} W_{t\mu}\left(A_{\mu\nu} + Q_{t\mu\nu}\right) W_{t\nu} - 2\sum_\mu W_{t\mu} v_{t\mu}\right), \tag{35}$$

where $\mathbf{A} \equiv \mathrm{diag}(\alpha_1, \alpha_2 \ldots \alpha_D)$ and we have defined:

$$Q_{t\mu\nu} = \sum_s M_{st}\Psi_{st}^{-1}\,\mathbb{E}\left[z_{s\mu} z_{s\nu}\right], \qquad v_{t\mu} = \sum_s M_{st}\Psi_{st}^{-1}(m_{st} - m_t)\,\mathbb{E}[v_{s\mu}]. \tag{36}$$

The maximum a posteriori result for $W_{t\mu}$ is:

$$W_{t\mu} = \sum_\nu \left(\mathbf{A} + \mathbf{Q}_t\right)^{-1}_{\mu\nu}\, v_{t\nu}. \tag{37}$$

In the approximation $\mathbf{\Sigma}_s \to \bar{\mathbf{\Sigma}}$, this formula is unchanged but $\mathbf{Q}_t$ is S times as fast to calculate. The other M steps and the E step are not affected by the ARD prior. To determine α_μ, we re-exponentiate expression (35)[4] and integrate out $\mathbf{W}$ to obtain the evidence for $\mathbf{A}$:

$$P(\mathbf{n}|\mathbf{A}) \propto \prod_k \alpha_k^{T/2} \prod_t \int q(\mathbf{W}|t) \prod_\mu \mathrm{d}W_{t\mu}, \tag{38}$$

---

[4] This is the distribution on $\mathbf{W}$ that we would obtain from a mean-field variational factorization $q(\mathbf{z})q(\mathbf{W})$.

where:

$$q(\mathbf{W}|t) \equiv \exp\left(-\frac{1}{2}\sum_{\mu\nu} W_{t\mu}(A_{\mu\nu} + Q_{t\mu\nu})W_{t\nu} - \sum_{\mu} W_{t\mu}v_{t\mu}\right). \tag{39}$$

The ARD coefficients α are determined by maximizing the log evidence. That is, we set

$$\frac{\partial}{\partial\alpha_k}\ln P(\mathbf{n}|\mathbf{A}) = 0 \Rightarrow \frac{1}{2}\left(\frac{T}{\alpha_\mu} - \sum_t \left\langle W_{t\mu}^2\right\rangle\right) = 0 \Rightarrow \alpha_\mu = \frac{T}{\sum_t \left\langle W_{t\mu}^2\right\rangle}, \tag{40}$$

where the average $\left\langle W_{t\mu}^2\right\rangle$ is taken with respect to the density $q(\mathbf{W}|t)$. Completing the square, we find that $q(\mathbf{W}_{t\cdot})$ is Gaussian with covariance $(\mathbf{A}+\mathbf{Q}_t)^{-1}$ and mean $(\mathbf{A}+\mathbf{Q}_t)^{-1}\mathbf{v}_t$. Note that this mean is precisely the M step value for $q(\mathbf{W}|t)$, as we would hope! Thus we get:

$$\left\langle W_{t\mu}^2\right\rangle = W_{t\mu}^2 + (\mathbf{A}+\mathbf{Q}_t)^{-1}_{\mu\mu} \tag{41}$$

Let us now summarize these steps:

- E step: $\mathbf{G}_s = \left(\mathbf{I} + \mathbf{W}^T\mathbf{M}_s\mathbf{\Psi}_s^{-1}\mathbf{W}\right)^{-1}$, $\mathbb{E}\left[\mathbf{z}_s\right] = \mathbf{G}_s\mathbf{W}^T\mathbf{\Psi}_s^{-1}(\mathbf{m}_s - \mathbf{m})$, $\mathbb{E}\left[\mathbf{z}_s\mathbf{z}_s^T\right] = \mathbf{G}_s + \mathbb{E}[\mathbf{z}_s]\mathbb{E}[\mathbf{z}_s]^T$. $\mathbf{G}_s$ and all the $\mathbb{E}[\mathbf{z}_s]$ are each $O(D^2TS)$. $\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^T]$ is $O(D^2S)$. The E step overall is $O(D^2TS)$.
- $\mathbf{m} = \left(\sum_s \mathbf{M}_s\mathbf{\Psi}_s^{-1}\right)^{-1}\sum_s\left[\mathbf{M}_s\mathbf{\Psi}_s^{-1}(\mathbf{m}_s - \mathbf{W}E[\mathbf{z}_s])\right]$ is $O(DTS)$.
- $B_{st} = \left(\mathbf{W}\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^T]\mathbf{W}^T\right)_{tt} + 2(m_t - m_{st})\left(\mathbf{W}\mathbb{E}[\mathbf{z}_s]\right)_t + (m_t - m_{st})^2$ is $O(D^2TS)$
- Solving Eq. (25) for each t is $O(TS)$ with a prefactor equal to the number of evaluations required to find a root (approximately $10 \sim 20$). As long as this number is less than D^2 this step is subleading.
- $\mathbf{Q}_t = \sum_s M_{st}\Psi_{st}^{-1}\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^T]$ is $O(D^2TS)$.
- $\mathbf{v}_t = \sum_s M_{st}\Psi_{st}^{-1}(m_{st} - m_t)E[\mathbf{z}_s]$ is $O(DTS)$.
- $\mathbf{W}_{t\cdot} = W_{t\mu} = \sum_\nu (\mathbf{A}+\mathbf{Q}_t)^{-1}_{\mu\nu}\, v_{t\nu}$ is $O(D^3T)$.
- $\left\langle W_{t\mu}^2\right\rangle = W_{t\mu}^2 + (\mathbf{A}+\mathbf{Q}_t)^{-1}_{\mu\mu}$ is $O(D^3T)$.
- $\alpha_\mu = T/\sum_t \left\langle W_{t\mu}^2\right\rangle$ is $O(DT)$.

The leading cost is a few terms of $O(D^2TS)$ flops, each with small prefactors, per iteration. Assuming a total prefactor of 10 and $T = 2\times10^5$, $D = 10$, $S = 500$ a full EM iteration costs 10^{11} flops in exact mode. On a single 1 GHz core (10^9 flops per second) this comes out to roughly 100 or 10 seconds.

We can apply the parameters learned from the panel of normals to single-sample calling, which requires the likelihood as a function of the copy numbers $\mathbf{c}$. Applying the same E step as above, the likelihood is

$$P(\mathbf{n}|\mathbf{c},\mathbf{W},\mathbf{m},\mathbf{\Psi}) \propto \prod_t \exp\left[-\frac{1}{2}M_{st}\Psi_{st}^{-1}\left(\ln\left(n_{st}/(d_s\mathcal{P}_{st})\right) - \ln(c_{st}) - m_t - \left(\mathbf{W}\mathbb{E}[\mathbf{z}_s]\right)_t\right)^2\right], \tag{42}$$

We have only kept factors that depends on $\mathbf{c}$ in the above likelihood. Note that this is factorized into independent likelihood terms for each target and is thus suitable for the emission model of an HMM. This likelihood is not Gaussian in $\mathbf{c}$, but it does not need to be for the Viterbi and forward-backward algorithms. Also, note that when $c_{st} = 0$ is the most likely solution, this must be incorporated in the mask matrix M_{st} in order to avoid ambiguous expressions due to the breakdown of the Laplace approximation used to replace the Poisson with a Gaussian.

1. GC bias correction

We can easily integrate sample-specific GC bias into this model. Let $f_s(\mathrm{GC})$ be the GC bias of GC content GC for sample s. Then this enters into the model as an additional multiplier to the Poisson parameter. That is, we replace $d_s c_{st} \to d_s c_{st} f_s(\mathrm{GC}_t)$, which affects the model learning and inference only via the definition of m_{st}. We can iteratively re-estimate the GC bias function f_s by regressing the bias not explained by the latent factors. That is, for each target the Poisson parameter is, ignoring GC effects,

$$c_{st} d_s \exp\left(\mathbf{W}_t \mathbb{E}[\mathbf{z}_s] + m_t\right) \tag{43}$$

and thus the ratio $n_{st} / \left[c_{st} d_s \exp\left(\mathbf{W}_t \mathbb{E}[\mathbf{z}_s] + m_t\right)\right]$ (with error bars of size $1/\sqrt{n}$ if we want to do a weighted regression) is an estimate of $f_s(\mathrm{GC}_t)$ that we can feed into our favorite regression model. This is more sophisticated than the standard approach of simply regressing n versus GC in that it seeks to explain with GC only the bias that cannot be explained by linear latent features.

2. PCA and the curse of small samples

Since this regularizer is quadratic in z, the Gaussian structure of the likelihood is preserved and the E step remains as simple as before. The only difference is the presence of an additional term in $\mathbf{G}_s^{-1}$:

$$\mathbf{G}_s^{-1} \to \mathbf{I} + \mathbf{W}^T \left[\mathbf{M}_s \mathbf{\Psi}_s^{-1} + \lambda \mathbf{F}\right] \mathbf{W}. \tag{44}$$

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

$$(\mathbf{W}^T \mathbf{F} \mathbf{W})_{\mu\nu} = \sum_{t=0}^{T-1} W_{\mu t}\, \mathrm{FFT}_t^{-1} \left[\sum_{k=0}^{T-1} \tilde{f}(k)\, \mathrm{FFT}_k[W_{t\nu}]\right]. \tag{45}$$

The M step equations for $\mathbf{\Psi}$ and $\mathbf{M}$ remain the same. For $\mathbf{W}$, we find:

$$\sum_{\nu} Q_{t\mu\nu} W_{t\nu} + \lambda \sum_{\nu, t'} Z_{\mu\nu} F_{tt'} W_{t'\nu} = v_{t\mu}, \tag{46}$$

where $\mathbf{Q}$ and $\mathbf{v}$ are defined as before and:

$$\mathbf{Z} = \sum_{s=1}^{S} \mathbb{E}[\mathbf{z}_s \mathbf{z}_s^T]. \tag{47}$$

As one would expect, the regularizer mixes different targets such that t is no longer a mere label in the stationarity condition for $\mathbf{W}$. The direct solution to Eq. (46) is impractical since it involves inverting a matrix of size $DT \times DT$.

Fortunately, the linear operator in question, $\mathbf{Q} + \lambda \mathbf{Z} \otimes \mathbf{F}$, is the sum of two sparse operators: $\mathbf{Q}$ is diagonal in the target space, and $\mathbf{A} \equiv \mathbf{Z} \otimes \mathbf{F}$ is diagonal in Fourier space ($\mathbf{Z}$ acts on the latent space, $\mathbf{F}$ acts on the target space). Both $\mathbf{Q}$ 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. (46) can be solved very efficiently using preconditioned iterative Krylov space solvers such as conjugate gradients or GMRES. A decent preconditioner for $\mathbf{A}$ can be constructed by taking a target average of $Q_{t\mu\nu}$:

$$\tilde{\mathbf{A}} \equiv \tilde{\mathbf{Q}} + \lambda \mathbf{Z} \otimes \mathbf{F}, \qquad \tilde{\mathbf{Q}} = \frac{1}{T} \sum_t \mathbf{Q}_t. \tag{48}$$

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

$$\left(\tilde{\mathbf{A}}^{-1}\mathbf{W}\right)_{t\mu} = \mathrm{FFT}_t^{-1} \left[\left(\tilde{\mathbf{Q}} + \lambda \tilde{f}(k) \mathbf{Z}\right)^{-1} \mathrm{FFT}_k[\mathbf{W}_t]\right]. \tag{49}$$

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), $\tilde{\mathbf{A}}^{-1}\mathbf{v}$ is an excellent approximate solution to Eq. (46) 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 $\mathcal{O}(D^3 T \log T)$.

Choice of λ — The regularizer "kicks in" when $\lambda \sim \Psi^{-1}$, as it can be inferred directly from Eq. (46). One may initially choose $\lambda \sim 1000\, \Psi^{-1}$ and progressively relax it as CNV calls stabilize.

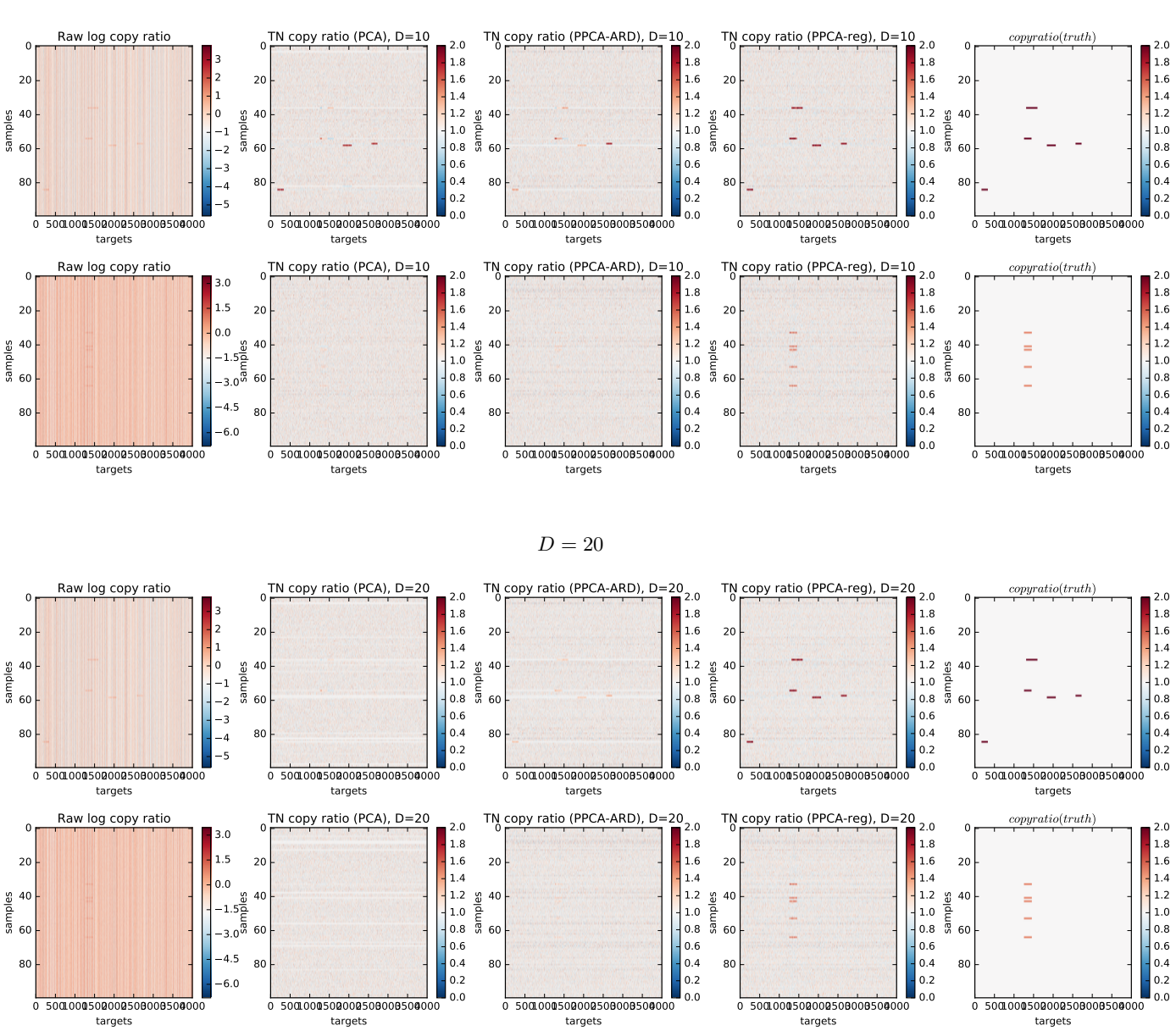

FIG. 2: Comparison of PCA with the probabilistic coverage model in different modes. Top two rows: $D = 10$; random events, correlated events. Bottom two rows: $D = 20$; random events, correlated events.

3. 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.

Figs. **??** and **??** 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.