Published in final edited form as:

Neural Comput. 2016 August ; 28(8): 1663–1693. doi:10.1162/NECO_a_00853.

A quasi-likelihood approach to non-negative matrix factorization

Karthik Devarajan and

Department of Biostatistics & Bioinformatics, Fox Chase Cancer Center, Temple University Health System, Philadelphia, PA 19111, karthik.devarajan@fccc.edu

Vincent C.K. Cheung

School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, vckc@cuhk.edu.hk

Abstract

A unified approach to non-negative matrix factorization based on the theory of generalized linear models is proposed. This approach embeds a variety of statistical models, including the exponential family, within a single theoretical framework and provides a unified view of such factorizations from the perspective of quasi-likelihood. Using this framework, a family of algorithms for handling signal-dependent noise is developed and its convergence proven using the Expectation-Maximization algorithm. In addition, a measure to evaluate the goodness-of-fit of the resulting factorization is described. The proposed methods allow modeling of non-linear effects via appropriate link functions and are illustrated using an application in biomedical signal processing.

Keywords

nonnegative matrix factorization; generalized linear models; quasi-likelihood; generalized Kullback-Leibler divergence; Expectation-Maximization algorithm; high-dimensional data; proportion of explained variation; signal-dependent noise; electromyography; motor module; muscle synergy

1 Introduction

Non-negative matrix factorization (NMF) is an unsupervised, parts-based learning paradigm, in which a high-dimensional nonnegative matrix V is decomposed into two non-negative matrices, W and H, such that $V = WH + \varepsilon$ where ε is the error matrix (Lee & Seung, 1999;2001). For a $p \times n$ matrix V consisting of n observations on each of p variables, each column of W defines a basis vector and each column of H represents an encoding of the corresponding observation. Each column of V can thus be approximated by a linear combination of the columns of W weighted by the elements of each column of H; and each row of V can be approximated by a linear combination of the rows of H weighted by the elements of each row of W. The number of basis vectors is determined by the rank k of the factorization. NMF has been increasingly applied to a variety of problems involving large-scale data that occur naturally on the non-negative scale. Some specific areas of application include computational biology, neuroscience, natural language processing, information

retrieval, biomedical signal processing and image analysis. A review of its applications can be found in Devarajan (2008).

The seminal work of Lee & Seung (2001) has spawned extensive research on this topic in the past decade from the perspectives of algorithm development and modeling. For instance, Hoyer (2004), Shahnaz et al. (2006), Pascual-Montano et al. (2006) and Berry et al. (2007) have extended the standard NMF algorithm to include sparseness constraints. Li et al. (2007), Lin (2007), Cichocki et al. (2007, 2009), Cichocki & Phan (2009), Wang & Li (2010), Févotte & Idier (2011), Phan & Cichocki (2011), Gillis & Glineur (2010, 2012), Zhou et al. (2012) and Guan et al. (2012, 2013) have proposed novel and efficient algorithms for NMF. Wang et al. (2006) extended the standard NMF algorithm to include uncertainty estimates and Ding et al. (2012) proposed a Bayesian non-parametric approach to NMF.

The standard formulation of the NMF problem typically assumes that the noise ε in the observed data follows a Gaussian model. Most of the above developments have focused heavily on this model; however, the observed data from diverse areas of application suggest a variety of scales and structures (Cheung & Tresch, 2005; Devarajan & Cheung, 2012; 2014). Although basic algorithms for certain non-Gaussian models exist, there has been virtually no development in terms of casting the NMF problem within a unifying and rigorous statistical and computational framework. In the high-dimensional setting the assumption of signal independence in noise has shown to be invalid in many applications, leading to a lack of robustness in the decomposed basis vectors W and encodings H, and poor reconstruction of the input matrix V (Cheung & Tresch, 2005; Devarajan & Cheung, 2014). Cheung & Tresch (2005) and Devarajan & Cheung (2012) extended NMFs to include members of the exponential family of distributions while Cichocki et al. (2006, 2008, 2009, 2011) developed generalized algorithms based on α- and β-divergences. Devarajan et al. (2005, 2008, 2015), Dhillon & Sra (2006) and Kompass (2007) have also proposed generalized divergence measures in this context. The work of Cichocki et al. (2009) remains a detailed reference on this subject.

In this paper, we generalize NMF to include *all* members of the exponential family of distributions within the framework of quasi-likelihood by exploiting signal-dependence in noise. Our flexible approach allows the likelihood that quantifies the divergence between the observed data matrix V and the reconstructed matrix WH to be defined based on a pre-specified statistical model determined by the structure of noise ε. Specifically, we develop a family of algorithms using the theory of generalized linear models that allows non-linear relationships between V and WH to be incorporated via link functions. Rigorous proofs of convergence and monotonicity of updates as well as a criterion for evaluating the goodness-of-fit of the resulting factorizations are provided. The methods are illustrated using an application in electromyography (EMG) studies; however, they are broadly applicable for dimension reduction of large-scale non-negative data arising from various other applications.

The rest of this paper is organized as follows. Section 2 provides the preliminary background and surveys existing work on NMF algorithms for handling signal-dependent noise. Section 3 develops the necessary theoretical framework using quasi-likelihood and proposes a unified family of NMF algorithms while section 4 describes a goodness-of-fit measure for

quantifying the factorizations from the proposed algorithms. In section 5, we illustrate our methods using experimental EMG data. The last section provides some concluding remarks.

2 Background

In many statistical and pattern recognition problems, the primary focus is on discriminating between two probability models F and G for a random prospect X that ranges over the space S. For an observation $X = x$, Bayes theorem relates the likelihood ratio to the prior and posterior odds in favor of F as $\log\frac{f(x)}{g(x)} = \log\frac{P(F|x)}{P(G|x)} - \log\frac{P(F)}{P(G)}$, where f and g are probability density functions and $P(\cdot)$ and $P(\cdot|x)$ denote the prior and posterior probabilities of the model. As the difference between the posterior and prior log-odds, the logarithm of the likelihood ratio $\log\left[\frac{f(x)}{g(x)}\right]$ quantifies the information in $X = x$ in favor of F against G.

If x is not given and no specific information is available on the whereabouts of x, other than $x \epsilon S$, then the mean observation per x from F for the discrimination information between F and G is

$$K(f,g) = \int_{\Re}\left(\log\frac{f(x)}{g(x)}\right)\mathrm{dF}(x), \quad (1)$$

given that F is absolutely continuous with respect to G, $F \lesssim G$. The discrimination information function in eqn. (1), known as Kullback-Leibler (KL) divergence, is a measure commonly used to compare two distributions, and was introduced in Kullback and Leibler (1951). This measure is nonnegative definite and is zero if and only if $f(x) = g(x)$ almost everywhere (Kullback, 1959; Ebrahimi and Soofi, 2004).

Generalized linear models (GLM) were introduced as a natural extension of classical linear models (Nelder & Wedderburn, 1972). GLMs include well-known models for handling various types of response data such as linear regression, Poisson regression models for count data, logit models for binary data and gamma models for non-negative data on the continuous scale. These different models share the key property of linearity that forms the core of the model-fitting algorithm in GLM. Consider the classical linear model $\mathbf{y} = X\beta + \epsilon$ where $\mathbf{y}$ is an n-vector of independent observations of a random variable Y with mean μ, X is a $n \times p$ matrix of observed covariates, β is a p-vector of parameters that need to be estimated from the data and ϵ denotes random error. The systematic component of the model is specified by $E(Y) = \mu = X\beta$, and the random component is specified by the assumption of independent and normally distributed errors ϵ with constant variance such that $E(Y) = \mu$ and $Var(Y) = \sigma^2$. If we denote the systematic and random components by η and μ, respectively, the link between these two components of the model is given by

$$\eta = g(\mu), \quad (2)$$

where $\eta = X\beta$ is the linear predictor and $g(.)$ is the link function that relates η to the expected value μ of $\mathbf{y}$. It turns out that in classical linear models, the mean and the linear predictor are identical i.e. the link $g(.)$ is the identity function. This is consistent with the assumption of normality and the range of possible values of η and μ. In general, the link function is a twice differentiable monotone function. Assuming that each independent observation y is a realization from a distribution in the exponential family, the log-likelihood for GLMs can be written as

$$l(\theta;y)=\frac{\{\theta y - b(\theta)\}}{a(\phi)}+c(y,\phi), \quad (3)$$

where $a(.)$, $b(.)$ and $c(.)$ are specific functions, and ϕ and θ are parameters (McCullaugh & Nelder, 1983). Using eqn. (3), it can be shown that $E(Y) = \mu = b'(\theta)$ and $Var(Y) = \sigma^2 = b''(\theta)a(\phi)$. Without loss of generality, we shall assume that ϕ is known and $a(\phi) \equiv 1$. Therefore, θ represents the canonical parameter of this exponential family. The link function in eqn. (2) is known as the canonical link if $\theta = \eta = g(\mu)$. For a given distribution, the canonical link has a sufficient statistic $X'\mathbf{y}$ whose dimension is equal to that of β in the linear predictor $\eta = X\beta$. Using the expression for μ written in terms of θ above, we can re-write the link function as $g(\mu) = g(b'(\theta)) = (gb')(\theta)$. For canonical links, $(gb')(\theta)$ is the identity function such that $\theta = \eta$ (McDonald & Diamond, 1990). In the case of Gaussian and Poisson distributions, the canonical links are, respectively, the identity and log links. A list of appropriate link functions for other members of the exponential family is provided in McCullaugh & Nelder (1983).

2.1 Existing work

The first step in obtaining an approximate factorization for V is to define cost functions that measure the divergence between the observed matrix V and the product of the factored matrices WH for a given rank r. We can express this in the form of a bi-linear model as $V = WH + \varepsilon$ where ε represents noise. Various cost functions have been proposed and utilized in the literature for assessing this factorization. As evidenced in the following section, these cost functions are typically derived from KL divergence (1) or its generalization based on an assumed (sometimes implicitly) statistical model for the data generating mechanism (or noise ε). One such metric is the Euclidean distance, $K(V\|WH) = \|V - WH\|^2$, which is based on the Gaussian likelihood and was proposed by Lee & Seung (2001). This quantity can be recognized as the KL divergence between V and WH for the Gaussian model (Devarajan et al., 2006; 2011; 2015). Lee & Seung (2001) also proposed a cost function based on the Poisson likelihood which they termed “KL” divergence. *It should be noted that we use the term KL divergence in a much broader context in this paper, one that is defined by eqn. (1) as the divergence between V and WH for any given statistical model specified by the signal-dependence in noise*.

Given the matrix V and factorization rank r, the goal is to find nonnegative matrices W and H such that $V \approx WH$. This is equivalent to minimizing the cost function $K(V\|WH)$ above with respect to W and H, subject to the constraints $W, H \geq 0$. Lee & Seung (2001) derived

multiplicative update rules for W and H based on random initial values. Furthermore, they showed that $K(V||WH)$ is non-increasing under these updates and that they are invariant under these updates if and only if W and H are at a stationary point of the divergence. Monotonicity of updates is theoretically established through the use of an auxiliary function similar to the one used in the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Due to the non-negativity requirement on the input matrix V and starting values for W and H, the multiplicative updates ensure non-negativity of the final converged solution for W and H.

Algorithms based on various generalized divergences have emerged since the original work of Lee & Seung (2001). Cheung & Tresch (2005) proposed heuristic algorithms for the exponential family of models and provided multiplicative update rules for W and H. In independent work, Cichocki et al. (2006) proposed heuristic algorithms for NMF based on the generalized β-divergence. Févotte & Idier (2011) proposed rigorous Majorization-Minimization (MM) and Majorization-Equalization (ME) algorithms for β-divergence that enabled monotonicity of updates for W and H to be theoretically established under different conditions. Furthermore, Cichocki et al. (2008, 2009, 2011) developed generalized algorithms based on α-divergence and proved monotonicity of updates using the EM algorithm. Devarajan et al. (2005, 2006, 2008, 2015) formulated a generalized approach to NMF based on the power divergence (PD) family of statistics that included various well-known divergence measures as special cases. Other generalized divergence measures have also been proposed (Dhillon & Sra, 2005; Kompass, 2007). Recently, Devarajan & Cheung (2014) outlined various algorithms for signal-dependent noise for the generalized inverse Gaussian family of models. From the perspective of statistical modeling, the literature is sparse and very little work has been done in terms of casting the NMF problem within a unifying and rigorous statistical framework.

3 A Quasi-likelihood Approach to NMF

The introduction of the quasi-likelihood (QL) approach significantly broadened the scope and applicability of GLMs by allowing a weaker assumption on the random component of the model (Wedderburn, 1974). This assumption specified only the mean-variance relationship rather than the underlying model itself. This idea was further extended by Nelder & Pregibon (1987) and it permitted the comparison of variance functions, linear predictors and link functions. In particular, the requirement that the variance function be known can be relaxed by embedding the variance function in a family of functions indexed by an unknown parameter α such that $Var(Y) = \phi\Sigma_\alpha(\mu)$ where $\phi > 0$ is the dispersion parameter in eqn. (4) below. For a single observation y, the deviance between y and its mean μ, $D_\alpha(y||\mu)$, can be written in terms of the *quasi-log-likelihood*, $Q(\mu|y)$, as

$$D_\alpha(y\|\mu) = -2\phi Q(\mu\|y) = -2\int_y^\mu \frac{y-u}{\sum_\alpha(u)}\mathrm{du}. \quad (4)$$

If $\Sigma_\alpha(\mu)=\mu^\alpha$, then

$$D_\alpha(y\|\mu)=\frac{2\{y^{2-\alpha}-(2-\alpha)y\mu^{1-\alpha}+(1-\alpha)\mu^{2-\alpha}\}}{(1-\alpha)(2-\alpha)},\alpha\in\Re\backslash\{1,2\}. \quad (5)$$

The power variance function used in eqn. (5) corresponds to an important class of exponential dispersion models (Jorgensen, 1987; Yilmaz & Cemgil, 2012). Special cases of the quantity in eqn. (5) include the Gaussian ($\alpha = 0$), Poisson ($\alpha \to 1$), Gamma ($\alpha \to 2$) and inverse Gaussian ($\alpha = 3$) models. It also includes the compound Poisson (CP) model when $1 < \alpha < 2$, the extreme stable (ES) distributions when $\alpha \leq 0$ and the positive stable (PS) distributions for $\alpha > 2$. The class of CP models is continuous for $y > 0$ but allows exact zeros. When $\alpha \to 1$ and $\phi > 0$ ($\phi \neq 1$), eqn. (5) corresponds to the Quasi-Poisson (QP) model that can be used to model over-dispersion ($\phi > 1$) or under-dispersion ($0 < \phi < 1$). It is well-known that an exponential family exists for $\alpha \leq 0$ and $\alpha \geq 1$ (Tweedie, 1981; Jorgensen, 1987). Using eqn. (1), the QL in eqn. (5) has an information-theoretic interpretation as the *generalized KL divergence, or discrimination information, of order* α between y and μ.

Divergence measures found in the NMF literature are special cases of $D_\alpha(y\|\mu)$ or related to it via variable transformations. These include β-divergence (Cichocki et al., 2006), α-divergence (Kompass, 2007; Cichocki et al., 2008) and Renyi divergence of order γ obtained using the PD family of statistics (Devarajan et al., 2005, 2008, 2015). However, none of them recognizes the formulation based on the mean-variance relationship in eqn. (5) or on the link function in eqn. (2). Let $\boldsymbol{\eta}$, $\boldsymbol{\mu}$ and θ represent matrix versions of the vector parameters η, μ and θ, respectively. A cost function can be written using $D_\alpha(y\|\mu)$ by expressing it in terms of V and WH for a particular choice of the link function $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = WH$ and by summing over the np observations in the input matrix V as follows:

$$D_\alpha(V\|g^{-1}(WH))=\sum_{i=1}^{p}\sum_{j=1}^{n}\frac{2\{V_{ij}^{2-\alpha}-(2-\alpha)V_{ij}[g^{-1}((WH)_{ij})]^{1-\alpha}+(1-\alpha)[g^{-1}((WH)_{ij})]^{2-\alpha}\}}{(1-\alpha)(2-\alpha)}, \quad (6)$$

where $\alpha \in \Re\backslash\{1, 2\}$. Using eqn. (6), β-divergence is obtained by setting $\alpha = 2 - \beta$ for the identity link function $g(\boldsymbol{\mu}) = \boldsymbol{\mu} = WH$ (up to a scale factor 0.5) and hence can be viewed as a special case of generalized KL divergence of order β. Writing the linear model $\upsilon = Wh + \varepsilon$ for the j^{th} column $\upsilon = (\upsilon_{1j}, \upsilon_{2j}, \ldots, \upsilon_{pj})'$ of V and using the link $\eta = g(\mu) = Wh$, the score equation derived from eqn. (6) has the form of a linear weighted least squares equation. Ignoring the term $2/(1 - \alpha)(2 - \alpha)$ and summing over the p observations, the score equation for a rank r factorization is given by

$$\frac{\partial D_\alpha}{\partial h_a}=\sum_{i=1}^{p}(v_{\mathrm{ij}}-\mu_{\mathrm{ij}})\left(\frac{d\theta}{d\eta}\right)w_{\mathrm{ia}},w_{\mathrm{ia}}\geq 0,h_a\geq 0,a=1,2,\ldots,r,\forall i,j, \quad (7)$$

with weight $\frac{d\theta}{d\eta}=\left(\frac{d\eta}{d\mu}\right)^{-1}[\sum_\alpha(\mu)]^{-1}$. For identity links, $\frac{d\theta}{d\eta}=[\sum_\alpha(\mu)]^{-1}$ and for canonical links $\frac{d\theta}{d\eta}=1$.

3.1 NMF as an alternating GLM algorithm

Consider a rank r decomposition of the $p \times n$ non-negative matrix V into $W_{p\times r}$ and $H_{r\times n}$, starting with random initial values for W and H. We can specify a linear model $\upsilon = Wh + \varepsilon$ for each column V where υ and h are column vectors of lengths p and r, respectively. Using the Gaussian framework for illustration, the linear predictor for this model is $\eta=g(\mu)=\mu=\mathrm{Wh}=\sum_{a=1}^{r}h_a w_{\mathrm{ia}}$ where the identity link is also the canonical link. For this model, since $\alpha = 0$ eqn. (6) reduces to

$$K(v\|\mathrm{Wh})=\sum_{i=1}^{p}(v_i-\sum_a h_a W_{\mathrm{ia}})^2 \quad (8)$$

where w_{ia} is the element of W at row i, column a. Here, W is known and h needs to be estimated so as to minimize the cost function in eqn. (8) where both W and h are constrained to be non-negative. Similarly, we can specify an appropriate linear model for every row of V as $\upsilon = H'w' + \varepsilon$ where υ and w' are column vectors of lengths n and r, respectively. We can then repeat the above steps to first estimate w given H and υ, and then estimate H given W and υ. Alternating between columns and rows of V and simultaneously updating W and H at each iteration is equivalent to applying the non-negativity constrained alternating GLM algorithm based on the Gaussian model. This is analogous to, and a generalization of, the non-negativity constrained alternating least squares approaches described elsewhere (Langville et al, Preprint; Kim & Park, 2006).

Generalizing this idea, minimization of the cost function (6) can be viewed as non-negativity constrained alternating minimization based on the GLM (ncAGLM) algorithm. Each side of this minimization is a convex problem that can be interpreted as a projection with respect to the divergence in eqn. (6) for a particular choice of α and link function $g(.)$. This cost function is alternatively minimized with respect to its two arguments W and H, each time estimating one argument while keeping the other fixed. Algorithms based on the Gaussian and Poisson likelihoods originally proposed by Lee & Seung (2001) can be viewed as instances of the ncAGLM algorithm using the identity link function (Devarajan & Cheung, 2012). Existing methods based on heuristic, EM and MM updates can all be viewed as special cases of this approach limited to use of the divergence in eqn. (6) and identity link. In the following sections, we show that this ncAGLM algorithm can be extended to not only

embed all members of the exponential family but also include canonical and other link functions thereby resulting in unified and efficient algorithms for NMF.

3.2 Canonical links

The link function connects the systematic component, $\boldsymbol{\eta} = WH$, to the mean of the data distribution. For a given distribution, there are many link functions that are acceptable in practice. However, the significance of the canonical link in GLMs cannot be overstated due to its desirable statistical properties. The predicted mean, $\boldsymbol{\mu}$, may not necessarily be mathematically the same as the data distribution's canonical location parameter. The canonical link $g(.)$ is the function that connects $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ such that $g(\boldsymbol{\mu}) = \boldsymbol{\theta}$, and has the form $g = b^{-1}$. The primary advantage of this is that the canonical link has a minimal sufficient statistic $W'v$ for h (and similarly Hv for w). In this case, it is well-known that the Newton-Raphson and Fisher scoring methods for maximum likelihood estimation coincide. The canonical link would ensure that the mean $\boldsymbol{\mu}$ is within the range of the observed data in V and that the residuals sum to zero. Furthermore, use of the canonical link provides an interpretation for algorithm-specific measures of goodness-of-fit such as the fraction of explained variation (R^2). For a rank r factorization based on the canonical link, R^2 measures the proportionate reduction in uncertainty due to the inclusion of W and H and, therefore, can be interpreted in terms of information content of the data. Such an interpretation is not provided by non-canonical links (for more details, see Cameron & Windmeijer, 1997 and §4).

The generalized form of the canonical link for the family of models represented by (6) can be written as $\boldsymbol{\eta}=g(\boldsymbol{\mu})=\frac{\boldsymbol{\mu}^{1-\alpha}}{1-\alpha}=\mathrm{WH}$ where the exponent is computed element-wise. Our choice of link function is motivated by the canonical link; however, in order to accommodate non-negativity restrictions on W, H and $\boldsymbol{\eta} = WH$, we propose the use of inverse power link,

$$\boldsymbol{\eta}=g(\boldsymbol{\mu})=\boldsymbol{\mu}^{1-\alpha}=\mathrm{WH}, \quad (9)$$

which leverages some of the attractive properties of canonical links. This family of link functions is useful for handling many forms of skewed distributions that are encountered in practice. The inverse polynomial response surfaces originally described in Nelder (1966, 1991) are included in these links as special cases. In a recent neuroscience review article, Buzsáki & Mizuseki (2014) demonstrated that most physiological and anatomical features of the brain are characterized by heavily skewed distributions thereby suggesting their fundamental importance in the structural and functional organization of the brain. Although each parameter vector $\mu = g(\mu) = Wh$ lies in a linear subspace, it corresponds to a nonlinear surface in the space of data (V) as suggested by the link function $g(.)$. The minimization of the cost function in eqn. (6) is a search for a lower dimensional nonnegative basis matrix W which defines a surface $\mathscr{S}(W)=\{g^{-1}(\mathrm{Wh})|h\in\Re_+^r\}$ that is close to the data points in V. The optimizer of W satisfies $W = \arg\min_W \Sigma_j \min_{s\in\mathscr{S}} D_\alpha(v_j \| \mathscr{S})$. The Gaussian model is a special case for which the identity link is the canonical link and $\mathscr{S}(W)$ is the hyperplane with W as its basis. For factorization of V using a particular statistical model, the choice of

link determines the weight $\frac{d\theta}{d\eta}$ and hence the form of the score function. The link function plays a significant role in the decomposition itself resulting in structurally different components and provides a variety of interpretations. It depends on the hypothesized or observed mean-variance relationship and specifies a particular nonlinear relationship between the systematic and random components of the model, and determines the form of the cost function. Table 1 lists selected variance functions, link functions and weights $\frac{d\theta}{d\eta}$ corresponding to some well-known statistical models that are applicable in NMF. It should be noted that other combinations of models and link functions are also possible and could be application dependent (see Remark 2 and §5 for more discussion).

3.3 A unified family of algorithms for NMF

Using the inverse power link in eqn. (9), the divergence in eqn. (6) can be written in terms of V and $g^{-1}(WH)$ as (ignoring the term $2/(1-\alpha)(2-\alpha)$)

$$D_\alpha(V\|g^{-1}(\mathrm{WH})) = \begin{cases} \sum_{i,j}\{V_{ij}^{2-\alpha} - (2-\alpha)V_{ij}[(\mathrm{WH})_{ij}] + (1-\alpha)[(\mathrm{WH})_{ij}]^{(\frac{2-\alpha}{1-\alpha})}\}, \\ \qquad\qquad \alpha \in (-\infty, 1) \cup (2, \infty) \\ \sum_{i,j}\{-V_{ij}^{2-\alpha} + (2-\alpha)V_{ij}[(\mathrm{WH})_{ij}] - (1-\alpha)[(\mathrm{WH})_{ij}]^{(\frac{2-\alpha}{1-\alpha})}\}, \\ \qquad\qquad 1<\alpha<2, \\ \sum_{i,j}\{-\log(\mathrm{WH})_{ij} - \log V_{ij} + V_{ij}(\mathrm{WH})_{ij} - 1\}, \alpha=2. \end{cases} \tag{10}$$

From here on, the divergence in eqn. (10) will be referred to as *GenKL* and the family of NMF algorithms based on it will be called the *GenKL* algorithms. For a single observation $y = V_{ij}$ with mean $\mu = g^{-1}(\eta) = g^{-1}((WH)_{ij})$ the well-known β-divergence, obtained using the identity link in eqn. (6), is convex in $\eta = (WH)_{ij} = g^{-1}((WH)_{ij}) = \mu$ only when $\alpha \in [1, 2]$; and for values of α outside this range, it can be expressed as the sum of convex, concave and constant functions (Févotte & Idier, 2011). In contrast, the divergence in eqn. (10) based on the inverse power link is convex in η for any α on the entire real line except $\alpha = 1$. This is evident from its second derivative with respect to $\eta = (\mathrm{WH})_{ij}$, $\frac{1}{(\alpha-1)^2}\eta^{\frac{\alpha}{1-\alpha}}$, and is illustrated in Figure 1. This result plays a significant role in the development of a unified family of algorithms for NMF via a generalization of the ncAGLM algorithm, as evidenced by Theorem 1 below. This family of algorithms generalizes update rules for the exponential family of models based on the inverse power link when $\alpha \leq 0$ or $\alpha > 1$. From Theorem 1, these updates are seen to be significantly different from those of existing algorithms.

Theorem 1—*For* $\alpha \in \Re\backslash\{1\}$, *GenKL divergence,* $D_\alpha(V\|g^{-1}(WH))$, *in eqn. (10) is non-increasing under the following update rules for H and W:*

$$H_{\mathrm{aj}}^{t+1}=H_{\mathrm{aj}}^{t}\left(\frac{\sum_i(\sum_b W_{\mathrm{ib}}H_{\mathrm{bj}}^{t})^{1/(1-\alpha)}W_{\mathrm{ia}}}{\sum_i W_{\mathrm{ia}}V_{\mathrm{ij}}}\right)^{\alpha-1} \quad (11)$$

and

$$W_{\mathrm{ia}}^{t+1}=W_{\mathrm{ia}}^{t}\left(\frac{\sum_j(\sum_b W_{\mathrm{ib}}^{t}H_{\mathrm{bj}})^{1/(1-\alpha)}H_{\mathrm{aj}}}{\sum_j H_{\mathrm{aj}}V_{\mathrm{ij}}}\right)^{\alpha-1}, \quad (12)$$

where H and W are non-negative. This measure is invariant under these updates if and only if H and W are at a stationary point of the divergence.

Proof—We provide a more general and rigorous proof of the monotonicity of updates based on splitting the domain $\Re\backslash\{1\}$ of the parameter α into four disjoint regions and considering each separately. First, we derive the update for H and prove its monotonicity for $\alpha < 1$. Then we show how similar arguments can be used to prove the result for $\alpha > 2$ and $1 < \alpha < 2$. The case $\alpha = 2$ is considered separately. In each case, we will use the convexity of x^{ν} where $\nu=\frac{2-\alpha}{1-\alpha}$ for a particular range of α. The update rules obtained under all cases, however, are the same.

We will make use of an auxiliary function similar to the one used in the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Lee & Seung, 2001; Devarajan et al., 2015). Note that for h real, $G(h, h')$ is an auxiliary function for $F(h)$ if $G(h, h') \geq F(h)$ and $G(h, h) = F(h)$ where G and F are scalar valued functions. Also, if G is an auxiliary function, then F is non-increasing under the update $h^{t+1}=\arg\min_h G(h,h^t)$. Using the first equation in (10), we define

$$F(H_{\mathrm{aj}})=\sum_i V_{\mathrm{ij}}^{2-\alpha}-(2-\alpha)\sum_i\left\{V_{\mathrm{ij}}\left(\sum_a W_{\mathrm{ia}}H_{\mathrm{aj}}\right)\right\}+(1-\alpha)\sum_i\left[\sum_a W_{\mathrm{ia}}H_{\mathrm{aj}}\right]^{\left(\frac{2-\alpha}{1-\alpha}\right)}, \quad (13)$$

where H_{aj} denotes the aj^{th} entry of H. Then the auxiliary function for $F(H_{aj})$ is

$$G(H_{\mathrm{aj}},H_{\mathrm{aj}}^{t})=\sum_i V_{\mathrm{ij}}^{2-\alpha}-(2-\alpha)\sum_i\left\{V_{\mathrm{ij}}\left(\sum_a W_{\mathrm{ia}}H_{\mathrm{aj}}\right)\right\}+(1-\alpha)\sum_{\mathrm{ia}}\left\{(W_{\mathrm{ia}}H_{\mathrm{aj}})^{\left(\frac{2-\alpha}{1-\alpha}\right)}\left(\frac{W_{\mathrm{ia}}H_{\mathrm{aj}}^{t}}{\sum_b W_{\mathrm{ib}}H_{\mathrm{bj}}^{t}}\right)^{1/(\alpha-1)}\right\}. \quad (14)$$

It is straightforward to show that $G(H_{aj}, H_{aj}) = F(H_{aj})$. To show that $G(H_{\mathrm{aj}}, H^t_{\mathrm{aj}}) \geq F(H_{\mathrm{aj}})$, we use the convexity of $x^{(\frac{2-\alpha}{1-\alpha})}$ for $\alpha < 1$ and the fact that for any convex function $f, f\left(\sum_{i=1}^{n} r_i x_i\right) \leq \sum_{i=1}^{n} r_i f(x_i)$ for rational nonnegative numbers $r_1, \cdots, r_n$ such that $\sum_{i=1}^{n} r_i = 1$.

Note that if $\alpha < 1$, then $\nu = \frac{2-\alpha}{1-\alpha} \geq 1$ and, hence, x^{ν} is convex. Thus, we obtain

$$\left(\sum_a W_{\mathrm{ia}} H_{\mathrm{aj}}\right)^{(\frac{2-\alpha}{1-\alpha})} \leq \sum_a \gamma_a \left(\frac{W_{\mathrm{ia}} H_{\mathrm{aj}}}{\gamma_a}\right)^{(\frac{2-\alpha}{1-\alpha})} = \sum_a \left\{ (W_{\mathrm{ia}} H_{\mathrm{aj}})^{(\frac{2-\alpha}{1-\alpha})} \left(\frac{W_{\mathrm{ia}} H^t_{\mathrm{aj}}}{\sum_b W_{\mathrm{ib}} H^t_{\mathrm{bj}}}\right)^{1/(\alpha-1)} \right\},$$

where $\gamma_a = \frac{W_{\mathrm{ia}} H^t_{\mathrm{aj}}}{\sum_b W_{\mathrm{ib}} H^t_{\mathrm{bj}}}$. From this inequality it follows that $F(H_{\mathrm{aj}}) \leq G(H_{\mathrm{aj}}, H^t_{\mathrm{aj}})$. The minimizer of $F(H_{aj})$ is obtained by solving

$$\frac{\mathrm{dG}(H_{\mathrm{aj}}, H^t_{\mathrm{aj}})}{dH_{\mathrm{aj}}} = -(2-\alpha)\sum_i W_{\mathrm{ia}} V_{\mathrm{ij}} + (2-\alpha)\sum_i \left\{ (W_{\mathrm{ia}} H_{\mathrm{aj}})^{1/(1-\alpha)} W_{\mathrm{ia}} \left(\frac{W_{\mathrm{ia}} H^t_{\mathrm{aj}}}{\sum_b W_{\mathrm{ib}} H^t_{\mathrm{bj}}}\right)^{1/(\alpha-1)} \right\} = 0.$$

The update rule for H thus takes the form given in (11).

Similarly, for $\alpha > 2$ we define the function $F(H_{aj})$ and its auxiliary function $G(H_{\mathrm{aj}}, H^t_{\mathrm{aj}})$ exactly as in equations (13) and (14), respectively. Observing that the last term on the right hand side of eqn. (14) is negative when $\alpha > 2$, we can use the convexity of $-x^{(\frac{2-\alpha}{1-\alpha})}$ to show that $F(H_{\mathrm{aj}}) \leq G(H_{\mathrm{aj}}, H^t_{\mathrm{aj}})$ and proceed to obtain the update rule for H as described above. Note that if $\alpha > 2$, then $0 \leq \nu \leq 1$ and, hence, $-x^{\nu}$ is convex. The update rule for this case is the same as that when $\alpha < 1$.

For $1 < \alpha < 2$, using the second equation in (10) we define

$$F(H_{\mathrm{aj}}) = \sum_i -V_{\mathrm{ij}}^{2-\alpha} + (2-\alpha)\sum_i \left\{ V_{\mathrm{ij}} \left(\sum_a W_{\mathrm{ia}} H_{\mathrm{aj}}\right)\right\} - (1-\alpha)\sum_i \left[\sum_a W_{\mathrm{ia}} H_{\mathrm{aj}}\right]^{(\frac{2-\alpha}{1-\alpha})},$$

and the auxiliary function for $F(H_{aj})$ as

$$G(H_{\mathrm{aj}}, H^t_{\mathrm{aj}}) = \sum_i -V_{\mathrm{ij}}^{2-\alpha} + (2-\alpha)\sum_i \left\{ V_{\mathrm{ij}} \left(\sum_a W_{\mathrm{ia}} H_{\mathrm{aj}}\right)\right\} - (1-\alpha)\sum_{\mathrm{ia}} \left\{ (W_{\mathrm{ia}} H_{\mathrm{aj}})^{(\frac{2-\alpha}{1-\alpha})} \left(\frac{W_{\mathrm{ia}} H^t_{\mathrm{aj}}}{\sum_b W_{\mathrm{ib}} H^t_{\mathrm{bj}}}\right)^{1/(\alpha-1)} \right\}.$$

It is easy to see that $G(H_{aj}, H_{aj}) = F(H_{aj})$. When $1 < \alpha < 2$, the third term on the right hand side of eqn. (15) is positive. Hence, by using the convexity of $x^{(\frac{2-\alpha}{1-\alpha})}$ for $1 < \alpha < 2$, we can

show that $F(H_{\mathrm{aj}}) \leq G(H_{\mathrm{aj}}, H_{\mathrm{aj}}^t)$ and proceed to obtain the update rule for H as described above. Note that if $1 < \alpha < 2$, then $\nu \leq 0$ and, hence, x^ν is convex. The update rule for this case is the same as that when $\alpha < 1$.

When $\alpha = 2$, we define

$$F(H_{\mathrm{aj}}) = \sum_i \left\{ -\log V_{\mathrm{ij}} - \log\left(\sum_a W_{\mathrm{ia}} H_{\mathrm{aj}}\right) + V_{\mathrm{ij}}\left(\sum_a W_{\mathrm{ia}} H_{\mathrm{aj}}\right) - 1 \right\}. \quad (16)$$

Its auxiliary function is

$$G(H_{\mathrm{aj}}, H_{\mathrm{aj}}^t) = \sum_i \left\{ V_{\mathrm{ij}}\left(\sum_a W_{\mathrm{ia}} H_{\mathrm{aj}}\right) - \log V_{\mathrm{ij}} - 1 - \sum_a \gamma_a \log\left(\frac{W_{\mathrm{ia}} H_{\mathrm{aj}}}{\gamma_a}\right) \right\}. \quad (17)$$

It is straightforward to show that $G(H_{aj}, H_{aj}) = F(H_{aj})$. Using the convexity of $-\log x$, we show that $F(H_{\mathrm{aj}}) \leq G(H_{\mathrm{aj}}, H_{\mathrm{aj}}^t)$. The minimizer of $F(H_{aj})$ is obtained by solving $\frac{\mathrm{dG}(H_{\mathrm{aj}}, H_{\mathrm{aj}}^t)}{dH_{\mathrm{aj}}} = 0$. Using (17), we get

$$\frac{\mathrm{dG}(H_{\mathrm{aj}}, H_{\mathrm{aj}}^t)}{dH_{\mathrm{aj}}} = \sum_i \left\{ V_{\mathrm{ij}} W_{\mathrm{ia}} - \left(\frac{W_{\mathrm{ia}} H_{\mathrm{aj}}^t}{H_{\mathrm{aj}} (\sum_b W_{\mathrm{ib}} H_{\mathrm{bj}}^t)}\right) \right\} = 0. \quad (18)$$

By using symmetry of the decomposition $V \sim WH$ and by reversing the arguments on W in each case above, one can easily obtain the update rule for W given in (12) in the same manner as H.

For a given α, we will start with random initial values for W and H and iterate until convergence, i.e, iterate until $|D_\alpha^{(i)}(V\|\mathrm{WH}) - D_\alpha^{(i-1)}(V\|\mathrm{WH})| < \delta$ where δ is a pre-specified threshold between 0 and 1 and i denotes the iteration number.

Remark 1—*GenKL* divergence in eqn. (10) includes members of the exponential family when $\alpha \leq 0$ or $\alpha > 1$. However, the unified algorithm in Theorem 1 is applicable for $\alpha \in \Re$ except when $\alpha = 1$, i.e. some models that are not members of the exponential family are also included. When $\alpha = 1$ (Poisson model), closed form updates for W and H cannot be derived using the inverse power link. An approximate solution can be obtained for values of α close to unity. The update rules for this model derived in Lee & Seung (2001) are implicitly based on the identity link.

Remark 2 (Non-canonical links)—In addition to the proposed family of algorithms utilizing the inverse power link, the QL approach can be used to further develop algorithms by incorporating certain non-canonical link functions $g(.)$ in eqn. (6) for a given statistical model determined by choice of α. Examples from the existing literature include heuristic,

EM, MM and ME-based algorithms for several members of the exponential family of models (Lee & Seung, 2001; Cheung & Tresch, 2005; Cichocki et al., 2006; Févotte & Idier, 2011). As noted in §3, many other generalized divergence measures are related to eqn. (6) via variable transformations and algorithms have been developed for these measures (Cichocki et al., 2006, 2008; Kompass, 2007; Devarajan & Ebrahimi, 2005, 2008; Devarajan et al., 2015). However, it is important to note that all these algorithms implicitly utilize the identity link, $g(\mu) = \mu$, in their formulations. In some applications, other link functions may be useful or even necessary. These currently available algorithms are only representative of the many combinations of link function and α that are possible using QL. In that sense, the QL approach has broader applicability and potential in the development of flexible, data-driven algorithms for NMF. Some such combinations may result in algorithms with additive, gradient-type update rules for W and H. By appropriately modifying the step-size, as in Cheung & Tresch (2005), these updates can be re-written as multiplicative updates. Further discussion on this topic is provided in §5.

Remark 3 (Estimation of ϕ)—It is evident from eqn. (4) that QL only depends multiplicatively on ϕ and that the divergence in eqn. (5) depends on y and μ and not on ϕ. Thus, for a single observation $y = V_{ij}$, ϕ does not affect estimation of $\hat{\mu} = g^{-1}((WH)_{ij})$. That is, ϕ does not play any role in the minimization of the divergence in eqns. (6) and (10) or in the derivation of update rules for W and H. Several methods have been suggested for the independent estimation of ϕ. Examples include the generalized Pearson χ^2 statistic and the mean deviance (Nelder & Pregibon, 1987; McCullaugh, 1983).

4 Measuring Goodness-of-fit

For a pre-specified order α and rank r, the update rules in (11) and (12) ensure monotonicity of updates for a given run based on random initial values for W and H. A common problem with NMF algorithms in general is that they may not necessarily converge to the same solution on each run, i.e., they are prone to the problem of local minima, thus requiring a given algorithm to be run using multiple random restarts. The factorization from the run resulting in the best reconstruction, quantified by minimum reconstruction error across multiple runs, can be used for measuring goodness-of-fit.

We propose a single, unified measure for this purpose based on algorithm-specific minimum reconstruction error, E. It quantifies the variation explained by the continuum of statistical models for signal dependent noise described by equation (10). For each pre-specified rank r the proportion of explained variation or empirical uncertainty, R^2, is dependent on the particular model, determined by the order α, used in the factorization. For *GenKL* algorithms, it is defined as

$$R^2=1-\frac{\min D_\alpha(V\|g^{-1}(\mathrm{WH}))}{D_\alpha(V\|\overline{V})}=\begin{cases}1-\frac{\sum_{i,j}\left\{V_{\mathrm{ij}}^{2-\alpha}-(2-\alpha)V_{\mathrm{ij}}[(\sum_{\alpha=1}^{r}W_{\mathrm{ia}}H_{\mathrm{aj}})]+(1-\alpha)[(\sum_{\alpha=1}^{r}W_{\mathrm{ia}}H_{\mathrm{aj}})]^{\left(\frac{2-\alpha}{1-\alpha}\right)}\right\}}{\sum_{i,j}\{V_{\mathrm{ij}}^{2-\alpha}-(2-\alpha)V_{\mathrm{ij}}\overline{V}^{1-\alpha}+(1-\alpha)\overline{V}^{2-\alpha}\}}, & \alpha\in\Re\backslash\{1\},\\ 1-\frac{\sum_{i,j}\{-\log(\sum_{\alpha=1}^{r}W_{\mathrm{ia}}H_{\mathrm{aj}})-\log V_{\mathrm{ij}}+V_{\mathrm{ij}}(\sum_{\alpha=1}^{r}W_{\mathrm{ia}}H_{\mathrm{aj}})-1\}}{\sum_{i,j}\{-\log V_{\mathrm{ij}}+\log\overline{V}+V_{\mathrm{ij}}/\overline{V}-1\}}, & \alpha=2,\end{cases}$$

(19)

where E is computed using the numerator of the right hand side of eqn. (19) for a particular rank r and order α. The quantity $\sum_{a=1}^{r}W_{\mathrm{ia}}H_{\mathrm{aj}}$ in this numerator is the $(i, j)^{th}$ entry of the reconstructed matrix - $g^{-1}((WH)_{ij})$ - for a given rank r. In the denominator, each entry of $g^{-1}(WH)$ is replaced by the grand mean of all entries of the input matrix V, $\overline{V}=\frac{1}{np}\left\{\sum_{i=1}^{p}\sum_{j=1}^{n}V_{\mathrm{ij}}\right\}$. When $\alpha = 0$, these quantities reduce to the residual sum of squares and total sum of squares, respectively, associated with the Gaussian model (Devarajan & Cheung, 2014). The calculation of R^2 is based on the principle that algorithm-specific minimum reconstruction error E (model deviance) quantifies the performance of the model, determined by the entries $g^{-1}((WH)_{ij})$, while in the absence of the factorization the best approximation of each entry is provided simply by the grand mean of all observations in the data (null deviance). This extends the definition of R^2 to the sequence of non-linear models indexed by order α. For non-Gaussian models based on the canonical link, R^2 measures the proportionate reduction in uncertainty (or variation) due to the inclusion of W and H and, therefore, can be interpreted in terms of information content of the data (see Cameron & Windmeijer, 1997; Devarajan & Cheung, 2014 for more details). However, for models based on non-canonical links (such as the identity link with the exception of the Gaussian model) R^2 measures the proportion of empirical uncertainty explained by the inclusion of W and H. Unlike other similar measures that can sometimes take negative values, R^2 falls in the interval [0, 1].

5 Application to EMG Data

Examples of using NMFs in statistical signal processing include the analysis of signals from neuronal activity, EMG and electroencephalography studies, sparse coding, speech recognition, imaging studies such as facial pattern recognition, video summarization, structural and functional magnetic resonance imaging, and computer assisted tomography (Lee & Seung, 1999; Hoyer, 2004; Cheung & Tresch, 2005; Devarajan et al., 2008; Lee et al., 2009; Anderson et al., 2014). Here, we present an example involving the extraction of muscle synergies - fundamental neural control modules for movement generation - from EMG data using NMF.

The recognition problem: Identifying muscle synergies from EMG data

In EMG studies, electrical signals are recorded from muscles that reflect how they are activated by the central nervous system (CNS) for a particular posture or movement. The EMG signal is a spatiotemporal summation of the motor action potentials traveling along the muscle fibers of the thousands of motor units in the recorded muscle. The high-frequency components of the EMGs reflect, in addition to the noise, contribution of these action potentials. It is well-known that EMG data exhibits signal-dependent noise (Harris & Wolpert, 1998; Cheung & Tresch, 2005). This signal dependence of noise amplitude originates partly from the fact that when a muscle is activated, the smaller motor units are always recruited before the larger motor units, an observation commonly known as “The Henneman’s size principle” (Henneman, 1957). To produce more force from the muscle, increasingly larger motor units are recruited; a larger fluctuation of force and EMG then results as large motor units are recruited and de-recruited. Modelling studies have demonstrated signal-dependent noise in force production if Henneman’s principle of orderly motor-unit recruitment holds true (Jones et al., 2002; Stein et al., 2005).

A much-studied question in neuroscience concerns how the motor system coordinates the activations of hundreds of skeletal muscles, representing hundreds of degrees of freedom to be controlled (Bernstein, 1967). The CNS likely circumvents the complexity of movement and postural control arising from high dimensionality by activating groups of muscles as individual neural modules, known as muscle synergies (Tresch et al., 2002; Giszter et al., 2007; Bizzi et al., 2008; Ting et al., 2015). NMF and other linear factorization algorithms facilitate the extraction of muscle synergies from the EMG data; the extracted vectors can naturally be interpreted as representations of basic, time-invariant neural modules of motor control whose existence has been demonstrated in physiological experiments (Bizzi & Cheung, 2013; Giszter, 2015). As modules of motor control, muscle synergies serve to reduce the search space of motor commands, reduce potential redundancy of motor commands for a given movement, and facilitate learning of new motor skills (Poggio and Bizzi, 2004).

In the context of locomotor behaviors such as swimming, walking and jumping, the muscle synergies identified from the EMGs can be interpreted as basic components of the central pattern generators, or spinal neuronal networks that can produce patterns of motor output even in the absence of sensory input (Grillner, 1985; Drew et al., 2008; Cheung et al., 2005; Dominici et al., 2011). In an experimental setting, the ideal factorization algorithm for use in muscle-synergy analysis should then identify locomotor muscle synergies that remain the same even after de-afferentation, or the experimental manipulation of depriving the spinal cord of sensory input by severing the dorsal nerve roots of an animal. Thus, in addition to R^2 and the rank (see below), the similarity between the muscle synergies extracted from EMGs collected before, and after, deafferentation is a quantity that reasonably measures the performance of an algorithm in EMG data.

An EMG dataset is typically presented as a $p \times n$ matrix V whose p rows correspond to different muscles, and the n columns to disjoint, sequentially sampled data integrated over a specific time interval. Thus, each column of V represents an activation vector in the muscle space at one time instance. The goal is to find a small number, r, of muscle synergies, each

defined as a nonnegative, time-invariant activation balance profile in the p-dimensional muscle space, by decomposing V into $W_{p\times r}$ (each column of which is a time-invariant muscle synergy) and $H_{r\times n}$ (each column contains activation coefficients for the r synergies in W for one time instance).

EMG studies have not traditionally accounted for signal-dependent noise in their formulations (Bizzi & Cheung, 2013; Saltiel et al., 2001; Tresch et al., 2006; Overduin et al., 2012). Cheung et al. (2005) presented EMG data obtained from four different motor behaviors of frogs - deafferented jump, intact jump, deafferented swim and intact swim - and later developed heuristic algorithms for signal-dependent noise (SDN) based on the gamma model (Cheung & Tresch, 2005). For more details on this data set, see Cheung et al. (2005) and Devarajan & Cheung (2014). Devarajan & Cheung (2014) developed several rigorous EM algorithms for SDN based on gamma and inverse Gaussian models using dual KL divergence and J-divergence and demonstrated the utility of these algorithms using this EMG data. These new algorithms outperformed the Gaussian as well as existing algorithms for signal-dependent noise (Lee & Seung, 2001; Cheung & Tresch, 2005; Févotte & Idier, 2011) and showed superior performance in terms of selecting the appropriate rank (r) based on Akaike Information Criterion (AIC) and the proportion of variation explained in the data (R^2). In particular, the algorithm based on symmetric J-divergence (J) for the gamma model - obtained as the sum of KL divergence (6) and its dual ($\alpha \rightarrow 2$ for the identity link) - showed the best overall performance. Furthermore, these studies highlighted the need for using the algorithm appropriate for the statistical model underlying the data based on its noise properties. Even under model mis-specification, the extracted synergies were seen to contain substantial information about the underlying structure as long as signal-dependence in noise was accounted for by the model in some fashion. In this paper, we demonstrate further improvement on these results by use of inverse power links and data-driven choice of α on this data.

Choice of α and link function

As the parameter α quantifies signal-dependence in noise, its choice is an important consideration in real data analysis. Using the approach in Cheung & Tresch (2005) and Devarajan & Cheung (2014), α is either determined *a priori* based on an assumed underlying statistical model for the data or is empirically estimated within acceptable limits from the data. It turns out that all existing methods for SDN assume a value of α based on an underlying statistical model (Cheung & Tresch, 2005; Cichocki et al., 2006; Févotte & Idier, 2011; Devarajan & Cheung, 2014). The algorithm based on J-divergence also suffers the same issue and is applicable only for the gamma model. On the other hand, the proposed approach is application-dependent and enables the use of a more precise, data-driven estimate of α rather than relying on an approximation that is determined by model assumptions. It should be emphasized that none of the existing methods employ canonical links and have been limited mostly to certain special cases of eqn. (6) that implicitly rely on the use of identity link functions. Thus, appropriate choice of α alone is not sufficient and consideration of the link function $g(.)$ is also necessary in any application. For a given α, canonical links offer significant advantages due to their desirable statistical properties as outlined in §3.2. Choice of the link function is intricately related to the hypothesized or

empirically determined signal-dependence in noise and specifies that the mean vector $\eta = g(\mu) = Wh$ (that approximates each column of the input matrix V) corresponds to a nonlinear surface in the space of data points in V. In essence, the link specifies how the underlying pattern-generation mechanism (WH) may be related to the observed EMG data, potentially even in a nonlinear manner. If we have a well-defined model of how EMG relates to WH, then we can specify an appropriate link function *a priori*. The link also provides different interpretations of the decomposition itself. This point is further addressed in the following section within the context of the EMG data where different choices of α were used in combination with various link functions. These corresponded to well-known statistical models (see Table 1) or were based on an estimate of the range of α using the data. This approach serves as a useful guide in the exploration of a high-dimensional data set.

Interpretation of the inverse power link for EMG data

The inverse power links for the gamma and inverse Gaussian models are, respectively, the inverse (reciprocal) and inverse squared transformations. The inverse linear and inverse quadratic response surfaces described in McCullaugh & Nelder (1983) and Nelder (1966, 1991) provide concrete examples of these links. For the gamma model, the link is specified as $\boldsymbol{\eta}=g(\boldsymbol{\mu})=\frac{1}{\boldsymbol{\mu}}=\mathrm{WH}$ and for the inverse Gaussian model it is specified as $\boldsymbol{\eta}=g(\boldsymbol{\mu})=\frac{1}{\boldsymbol{\mu}^2}=\mathrm{WH}$. In both cases, the requirement that $\boldsymbol{\mu} > 0$ works seamlessly with the nonnegativity requirement on W and H ensuring nonnegativity of $\boldsymbol{\eta}$. Similarly, for the family of models embedded in *GenKL* divergence (10) the link is specified by eqn. (9), and this includes the exponential family of models for $\alpha \leq 0$ and $\alpha > 1$. In the context of EMG data, use of the inverse link for the gamma model confers a linear relationship between the EMG signal and the reciprocal of the entries of W or H. The entries of H, under this link function, may contain information about inhibitory drives sent to the muscle synergies which are not well captured by current models of muscle-synergy combination (Tresch et al., 2006), but are likely functionally relevant to the physiology of spinal motor modules (Jordan, 1991). Even if muscular compositions of the synergies extracted from the inverse link correspond to those extracted by other algorithms, the interpretation of their activations captured in H from the inverse link perspective would be totally different, in that each synergy in W would be seen as normally being inhibited (higher values in the encodings H), and their expression relying on a dis-inhibition (a decrease in the encodings H). There has been physiological experience suggesting that motor pattern formation depends critically on the regulation of inhibitory drives. For instance, the Renshaw cells, which inhibit motoneuronal activities, are themselves modulated by other neurons within the spinal central pattern generators (Nishimaru et al., 2006). Recent data obtained using optogenetics have also demonstrated the existence of inhibitory interneurons in the spinal cord that can globally suppress the activations of multiple muscles (Caggiano et al., 2014); thus, it is entirely plausible that some muscle patterns can only be expressed via dis-inhibition of these inhibitory neurons.

Generalizing this interpretation of H using the notion of signal-dependence in noise, we postulate that (i) H obtained using the inverse power link contains information about inhibitory drives when both excitatory and inhibitory drives to each muscle synergy are

present, and (ii) W contains information about how muscles are dis-inhibited together as a group in a coordinated fashion. The exponent is determined by the mean-variance relationship. This formulation not only includes the inverse and inverse squared links but also encompasses the family of inverse power links with their attractive properties. This approach may provide novel insights into principles underlying the muscular compositions of muscle synergies.

Muscle synergies extracted by the proposed family of algorithms

The utility of NMF algorithms based on inverse power links was explored in a subset of the aforementioned frog EMG data. The performance of a variety of existing algorithms were compared with that of the proposed family of algorithms in terms of the proportion of variation explained (R^2) in the data at the rank (number of synergies, r) determined based on minimum AIC. For the range of models considered here, AIC was calculated for a given algorithm using the minimum reconstruction error E (based on the particular model and divergence used) following the approach outlined in Devarajan & Cheung (2014). Existing methods used in this comparison included that of Lee & Seung (2001), Cheung & Tresch (2005), Cichocki et al., (2006), Févotte & Idier (2011), Devarajan & Cheung (2014) and Cheung et al. (2015). Table 2 lists the various statistical models, algorithms and link functions used in the analyses and summarizes their performance on the frog motor behaviors. The proposed family of algorithms are appropriately identified in this table in order to facilitate comparisons. Our results suggest that *GenKL* algorithms utilizing inverse power links outperformed existing and recently proposed methods for SDN in identifying the appropriate, physiologically interpretable number of synergies. The form of the link itself is determined in part based on the nature of signal-dependence inherent in the noise. An increase in R^2, as compared with the Gaussian algorithm, was observed for *GenKL* algorithms based on the respective inverse power links relative to methods implicitly based on the identity link. In particular, the best overall performance was seen when $\alpha \in (2.42, 2.50)$. In these cases, the algorithms consistently identified the number of synergies to be between 3 and 5 for all four behaviors, numbers that have been established as physiologically interpretable using kinematic analysis (d'Avella et al., 2003; d'Avella and Bizzi, 2005), while at the same time achieving an R^2 of over 85%. These results not only corroborate our previous findings on the empirical mean-variance relationship in EMG signals (for details, see Devarajan & Cheung, 2014) but also demonstrate the advantages of our unified approach for the exponential family of models. Specifically, when $\alpha = 2.42$, the ranks for the intact and deafferented jump data sets were found to be 3, and the ranks for the intact and deafferented swim data sets, to be 4 (Table 2). These ranks are identical to the optimal ranks we identified for the same data sets using the AIC in our previous study (Devarajan & Cheung, 2014). To facilitate comparison with our previous results, we shall use results from the *GenKL* algorithm when $\alpha = 2.42$ in our subsequent analysis described below.

As noted earlier, the ideal NMF algorithm for muscle-synergy analysis should return similar synergies from experimental EMGs recorded before and after deafferentation, respectively. We quantified performance of different NMF algorithms by calculating the best-matching scalar product between the synergy vectors from the intact and deafferented data sets. For

jump, in all four frogs performance of the *GenKL* algorithm ($\alpha = 2.42$) was comparable to that of *J* in finding synergies that persisted after deafferentation, one of the best-performing NMF algorithms in our previous study (Devarajan & Cheung, 2014) (Figure 2A). For swim, in three out of four frogs *GenKL* performed comparably well or even better than *J* (Figure 2B). As an example, we show in Figure 3 the swim muscle synergies of frog 2 returned by the Gaussian (Figure 3A), *J* (Figure 3B), and *GenKL* (Figure 3C) algorithms, with Figure 3A-B being identical to Figure 5A-B in our previous study (Devarajan & Cheung, 2014). The four intact (Figure 3, black bars) and deafferented (white bars) synergy pairs extracted by the *GenKL* algorithm were even more similar to each other than those extracted by the Gaussian and *J* algorithms. Specifically, synergy 1 from *GenKL* (muscles AD, SM, ST, IP) corresponds to synergy 4 from *J*. While the latter pair of intact/deafferented synergies were similar only in the group of muscles activated but not in their activation balance, the former pair were identical in both the group of muscles activated and in their activation balance. Synergy 2 (AD, SM, VI, GA, VE) and synergy 3 (VI, RA, GA, VE) from *GenKL* appear to be fractionations of synergy 3 from Gaussian or *J*, while synergy 4 from *GenKL* (IP, VI, TA, PE, BI, SA, VE) corresponds roughly to a merging of synergies 1 and 2 from Gaussian or *J*. Overall, for this data set the *GenKL* algorithm outperformed the other two algorithms in identifying muscle synergies that remained invariant after deafferentation, suggesting that the *GenKL* synergies are likely closer to the true coordinative components that comprise the spinal locomotive central pattern generators.

We illustrate in Figure 4 how the *GenKL* synergies (*W*) and time-varying coefficients (*H*) can be combined to reconstruct the recorded EMGs, through the inverse power link, in a specific post-deafferentation swimming episode from frog 2. The EMG reconstructions (Figure 4, gray shadings) matched the experimental data (thick black lines) quite well in all muscles except RI, in which the reconstructions were mostly above the data, and BI, in which the reconstructions tended to be below the data. More importantly, the coefficient of each synergy was active when its corresponding muscles were not activated, illustrating how the coefficient of a synergy may be interpreted as the extent to which the synergy is inhibited. This example shows the feasibility of explaining EMG data by a framework that models motor outputs as groups of muscles that are dis-inhibited together in a coordinated fashion.

Summary and Conclusions

In summary, we have proposed a unified theoretical framework for NMF based on quasi-likelihood that represents a continuum of statistical models including all members of the exponential family. Using this framework, we developed a flexible generalization of NMF algorithms suitable for handling data with signal-dependent noise. A rigorous proof of monotonicity of updates has been provided and an algorithm-specific measure for quantifying the fraction of variation explained has been proposed. The proposed framework is the only one of its kind that encompasses the CP, QP, ES and PS families of models in addition to well-known members of the exponential family. Furthermore, this framework enables use of link functions that allow modeling of non-linear effects. One of the objectives of the proposed methods has been to provide improved algorithms based on the family of inverse power links for effectively extracting the underlying components in a variety of

studies involving biomedical signal processing. This is corroborated by numerical results presented on experimental EMG data in §5 where examples provide an intuitive interpretation of certain link functions useful for analyzing such data.

Acknowledgments

Research of K.D. was supported in part by NIH Grant P30 CA06927 and research of V. C. K. C. was supported by funds from The Chinese University of Hong Kong.

References

1. Anderson A, Douglas PK, Kerr WT, Haynes VS, Yuille AL, Xie J, Wu YN, Brown JA, Cohen MS. Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD. Neuroimage. 2014; 102:207–219. [PubMed: 24361664]
2. Bernstein, N. The coordination and regulation of movements. Oxford: Pergamon; 1967.
3. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis. 2007; 52:155–173.
4. Bizzi E, Cheung VCK, d'Avella A, Saltiel P, Tresch M. Combining modules for movement. Brain Research Reviews. 2008; 57:125–133. [PubMed: 18029291]
5. Bizzi E, Cheung VCK. The neural origin of muscle synergies. Frontiers in Computational Neuroscience. 2013; 7:51. [PubMed: 23641212]
6. Buzsáki G, Mizuseki K. The log-dynamic brain: how skewed distributions affect network operations. Nature Reviews Neuroscience. 2014; 15(4):264–278. [PubMed: 24569488]
7. Caggiano V, Sur M, Bizzi E. Rostro-Caudal inhibition of hindlimb movements in the spinal cord of mice. PLoS One. 2014; 9(6):e100865. [PubMed: 24963653]
8. Cameron AC, Windmeijer FAG. An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics. 1997; 77(2):329–342.
9. Cheung, VCK., Tresch, MC. Nonnegative matrix factorization algorithms modeling noise distributions within the exponential family; Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference; 2005. p. 4990-4993.
10. Cheung VCK, d'Avella A, Tresch MC, Bizzi E. Central and sensory contributions to the activation and organization of muscle synergies during natural motor behaviors. Journal of Neuroscience. 2005; 25(27):6419–6434. [PubMed: 16000633]
11. Cheung, VCK., Devarajan, K., Severini, G., Turolla, A., Bonato, P. Decomposing time series data by a non-negative matrix factorization algorithm with temporally constrained coefficients; Conference Proceedings of IEEE Engineering in Medicine and Biology Society 2015; 2015. p. 3496-3499.
12. Cichocki, A., Zdunek, R., Amari, S. Lecture Notes in Computer Science, Independent Component Analysis and Blind Signal Separation. Springer; 2006. Csiszar's Divergences for Non-negative Matrix Factorization: Family of New Algorithms; p. 32-39.LNCS-3889
13. Cichocki, A., Zdunek, R., Amari, S. Lecture Notes in Computer Science. Vol. 4666. Springer; 2007. Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization; p. 169-176.
14. Cichocki A, Lee H, Kim Y-D, Choi S. Non-negative matrix factorization with α-divergence. Pattern Recognition Letters. 2008; 29(9):1433–1440.
15. Cichocki, A., Zdunek, R., Phan, AH., Amari, S. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley; 2009.
16. Cichocki A, Phan HA. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. 2009; E92-A(3):708–721.

17. Cichocki A, Cruces S, Amari S. Generalized Alpha-Beta divergences and their application to robust nonnegative matrix factorization. Entropy. 2011; 13:134–170.
18. d'Avella A, Saltiel P, Bizzi E. Combinations of muscle synergies in the construction of a natural motor behavior. Nature Neuroscience. 2003; 6(3):300–308. [PubMed: 12563264]
19. d'Avella A, Bizzi E. Shared and specific muscle synergies in natural motor behaviors. Proceedings of the National Academy of Sciences, USA. 2005; 102(8):3076–3081.
20. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. 1977; 39:1–38.
21. Devarajan, K., Ebrahimi, N. Molecular pattern discovery using nonnegative matrix factorization based on Renyi's information measure; XII SCMA International Conference; 2-4 December 2005; Auburn, Alabama. 2005. http://atlas-conferences.com/c/a/q/t/98.htm
22. Devarajan, K. Proceedings of the Joint Statistical Meetings. Seattle; Washington: 2006. Nonnegative matrix factorization - A new paradigm for large-scale biological data analysis. CD-ROM
23. Devarajan K. Nonnegative matrix factorization - An analytical and interpretive tool in computational biology. PLoS Computational Biology. 2008; 4(7):E1000029. [PubMed: 18654623]
24. Devarajan K, Ebrahimi N. Class discovery via nonnegative matrix factorization. American Journal of Management and Mathematical Sciences. 2008; 28(34):457–467.
25. Devarajan, K. Problem Solving Handbook in Computational Biology and Bioinformatics. Springer; 2011. Matrix and Tensor Decompositions; p. 291-318.
26. Devarajan, K., Cheung, VCK. Joint Statistical Meetings. San Diego; California: 2012. On the relationship between non-negative matrix factorization and generalized linear modeling.
27. Devarajan K, Cheung VCK. On nonnegative matrix factorization algorithms for signal-dependent noise with application to electromyography data. Neural Computation. 2014; 26(6):1128–68. [PubMed: 24684448]
28. Devarajan K, Wang G, Ebrahimi N. A unified statistical approach to nonnegative matrix factorization and probabilistic latent semantic indexing. Machine Learning. 2015; 99(1):137–163. [PubMed: 25821345]
29. Dhillon, IS., Sra, S. Advances in Neural Information Processing Systems. Vol. 18. MIT Press; 2005. Generalized nonnegative matrix approximations with Bregman divergences.
30. Ding, N., Qi, Y., Xiang, R., Molloy, I., Li, N. Nonparametric Bayesian matrix factorization by Power-EP; Proceedings of the 13th International Conference on Artificial Intelligence and Statistics; 2012.
31. Dominici N, Ivanenko YP, Cappellini G, et al. Locomotor primitives in newborn babies and their development. Science. 2011; 334:997–999. [PubMed: 22096202]
32. Drew T, Kalaska J, Krouchev N. Muscle synergies during locomotion in the cat: a model for motor cortex control. Journal of Physiology. 2008; 586(Pt 5):1239–1245. [PubMed: 18202098]
33. Ebrahimi, N., Soofi, E. Information functions for Reliability. In: Soyer, R.Mazzuchi, TA., Singpurwalla, ND., editors. Mathematical Reliability, An Expository Perspective. Kluwer's International; 2004. p. 127-159.
34. Févotte C, Idier J. Algorithms for nonnegative matrix factorization with the β-divergence. Neural Computation. 2011; 23(9):2421–2456.
35. Gillis N, Glineur F. Using underapproximations for sparse nonnegative matrix factorization. Pattern Recognition. 2010; 43(4):1676–1687.
36. Gillis N, Glineur F. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. Neural Computation. 2012; 24(4):1085–1105. [PubMed: 22168561]
37. Giszter SF. Motor primitives - new data and future questions. Current Opinion in Neurobiology. 2015; 33:156–165. [PubMed: 25912883]
38. Giszter S, Patil V, Hart C. Primitives, premotor drives, and pattern generation: a combined computational and neuroethological perspective. Progress in Brain Research. 2007; 165:323–346. [PubMed: 17925255]

39. Grillner S. Neurological bases of rhythmic motor acts in vertebrates. Science. 1985; 228:143–149. [PubMed: 3975635]
40. Guan N, Tao D, Luo Z, Yuan B. NeNMF: An optimal gradient method for nonnegative matrix factorization. IEEE Transactions on Signal Processing. 2012; 60(6):2882–2898.
41. Guan N, Wei L, Luo Z, Tao D. Limited-memory fast gradient descent method for graph regularized non-negative matrix factorization. PLoS One. 2013; 8(10):e77162. [PubMed: 24204761]
42. Harris CM, Wolpert DM. Signal-dependent noise determines motor planning. Nature. 1998; 394:780–784. [PubMed: 9723616]
43. Henneman E. Relation between size of neurons and their susceptibility to discharge. Science. 1957; 126:1345–1347. [PubMed: 13495469]
44. Hoyer PO. Nonnegative matrix factorization with sparseness constraints. Journal of Machine Learning Research. 2004; 5:1457–1469.
45. Jones KE, Hamilton AF, Wolpert DM. Sources of signal-dependent noise during isometric force production. Journal of Neurophysiology. 2002; 88:1533–1544. [PubMed: 12205173]
46. Jordan, L. Brainstem and spinal cord mechanisms for the initiation of locomotion. In: Shimamura, M.Grillner, S., Edgerton, VR., editors. Neurological Basis of Human Locomotion. Tokyo: Japan Scientific Societies Press; 1991. p. 3-20.
47. Jorgensen B. Exponential dispersion models. Journal of the Royal Statistical Society-Series B. 1987; 49(2):127–162.
48. Kim, H., Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares; Proceedings of the IASTED International Conference on Computational and Systems Biology 95-100; 2006.
49. Kompass R. A generalized divergence measure for nonnegative matrix factorization. Neural Computation. 2007; 19:780–791. [PubMed: 17298233]
50. Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics. 1951; 22:79–86.
51. Kullback, S. Information Theory and Statistics. New York: Wiley; 1959.
52. Langville AN, Meyer CD, Albright R, Cox J, Duling D. Initializations, Algorithms, and Convergence for the Nonnegative Matrix Factorization. Preprint.
53. Lee DD, Seung SH. Learning the parts of objects by nonnegative matrix factorization. Nature. 1999; 401:788–791. [PubMed: 10548103]
54. Lee DD, Seung SH. Algorithms for nonnegative matrix factorization. Advances In Neural Information Processing Systems. 2001; 13:556–562.
55. Lee H, Cichocki A, Choi S. Kernel nonnegative matrix factorization for spectral EEG feature extraction. Neurocomputing. 2009; 72:3182–3190.
56. Li H, Adali T, Wang W, Emge D, Cichocki A. Nonnegative matrix factorization with orthogonality constraints and its application to Raman Spectroscopy. Journal of VLSI Signal Processing. 2007; 48:83–97.
57. Lin C-J. Projected gradient methods for non-negative matrix factorization. Neural Computation. 2007; 19:2756–2779. [PubMed: 17716011]
58. McCullaugh P. Quasi-likelihood functions. Annals of Statistics. 1983; 11(1):59–67.
59. McCullaugh, P., Nelder, JA. Generalized linear models. Chapman and Hall; London: 1983.
60. McDonald JW, Diamond ID. On the fitting of generalized linear models with nonnegativity parameter constraints. Biometrics. 1990; 46(1):201–206.
61. Nelder JA. Inverse polynomials, a useful group of multi-factor response functions. Biometrics. 1966:128–141.
62. Nelder JA. Generalized linear models for enzyme-kinetic data. Biometrics. 1991; 47:1605–1615. [PubMed: 1786334]
63. Nelder JA, Wedderburn RWM. Generalized linear models. Journal of the Royal Statistical Society, Series A. 1972; 135:370–384.
64. Nelder JA, Pregibon D. An extended quasi-likelihood function. Biometrika. 1987; 74(2):221–232.

65. Nishimaru H, Restrepo CE, Kiehn O. Activity of Renshaw cells during locomotor-like rhythmic activity in the isolated spinal cord of neonatal mice. Journal of Neuroscience. 2006; 26(20):5320–5328. [PubMed: 16707784]

66. Overduin SA, d'Avella A, Carmena JM, Bizzi E. Microstimulation activates a handful of muscle synergies. Neuron. 2012; 76(6):1071–7. [PubMed: 23259944]

67. Pascual-Montano A, Carazo JM, Kochi K, Lehmann D, Pascual-Marqui RD. Nonsmooth nonnegative matrix factorization (nsNMF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006; 28(3):403–415. [PubMed: 16526426]

68. Phan A-H, Cichocki A. Extended HALS algorithm for nonnegative Tucker decomposition and its applications for multiway analysis and classification. Neurocomputing. 2011; 74(11):1956–1969.

69. Poggio T, Bizzi E. Generalization in vision and motor control. Nature. 2004; 431(7010):768–774. [PubMed: 15483597]

70. Saltiel P, Wyler-Duda K, d'Avella A, Tresch MC, Bizzi E. Muscle synergies encoded within the spinal cord: evidence from focal intraspinal NMDA iontophoresis in the frog. Journal of Neurophysiology. 2001; 85(2):605–619. [PubMed: 11160497]

71. Shahnaz F, Berry M, Pauca VP, Plemmons RJ. Document clustering using nonnegative matrix factorization. Information Processing and Management: An International Journal. 2006; 42(2): 373–386.

72. Stein RB, Gossen ER, Jones KE. Neuronal variability: noise or part of the signal? Nature Reviews Neuroscience. 2005; 6:389–397. [PubMed: 15861181]

73. Ting LH, Chiel HJ, Trumbower RD, Allen JL, McKay JL, Hackney ME, Kesar TM. Neuromechanical Principles Underlying Movement Modularity and Their Implications for Rehabilitation. Neuron. 2015; 86(1):38–54. [PubMed: 25856485]

74. Tresch MC, Saltiel P, d'Avella A, Bizzi E. Coordination and localization in spinal motor systems. Brain Research Reviews. 2002; 40:66–79. [PubMed: 12589907]

75. Tresch MC, Cheung VCK, d'Avella A. Matrix factorization algorithms for the identification of muscle synergies: evaluation on simulated and experimental data sets. Journal of Neurophysiology. 2006; 95(4):2199–2212. [PubMed: 16394079]

76. Tweedie, MCK. An index which distinguishes between some important exponential families; Proceedings of the Indian Golden Jubilee International Conference on Statistics: Applications and New Directions; Calcutta, India. December 16-19, 1981; 1981. p. 579-604.

77. Yilmaz, YK., Cemgil, AT. Alpha/Beta divergences and Tweedie models. 2012. arXiv:1209.4280v1

78. Wang, F., Li, P. Efficient non-negative matrix factorization with random projections; Proceedings of The 10th SIAM International Conference on Data Mining; 2010. p. 281-292.

79. Wang G, Kossenkov AV, Ochs MF. LS-NMF: A modified nonnegative matrix factorization algorithm utilizing uncertainty estimates. BMC Bioinformatics. 2006; 7:175. [PubMed: 16569230]

80. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika. 1974; 61:439–447.

81. Zhou G, Cichocki A, Xie S. Fast Nonnegative Matrix/Tensor Factorization Based on Low-Rank Approximation. IEEE Transaction on Signal Processing. 2012; 60(6):2928–2940.

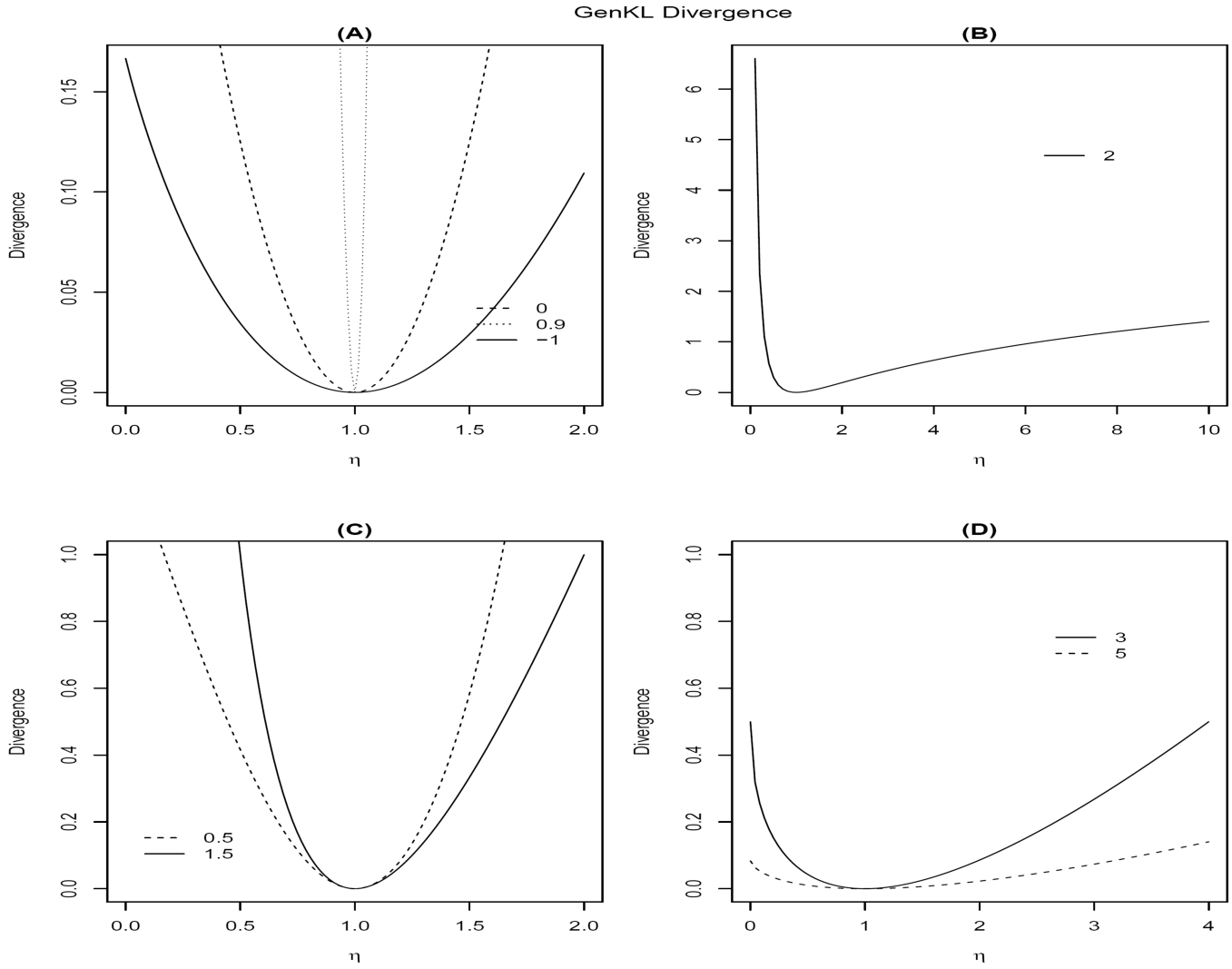

Figure 1.

(A)-(D), *GenKL* divergence, eqn. (10), plotted as a function of $\eta = g(\mu) = \mu^{1-\alpha} = (WH)_{ij}$ (inverse power link) for $V_{ij} = 1$ and various choices of α, illustrating its convexity. The values of α are indicated in the legend within each panel.

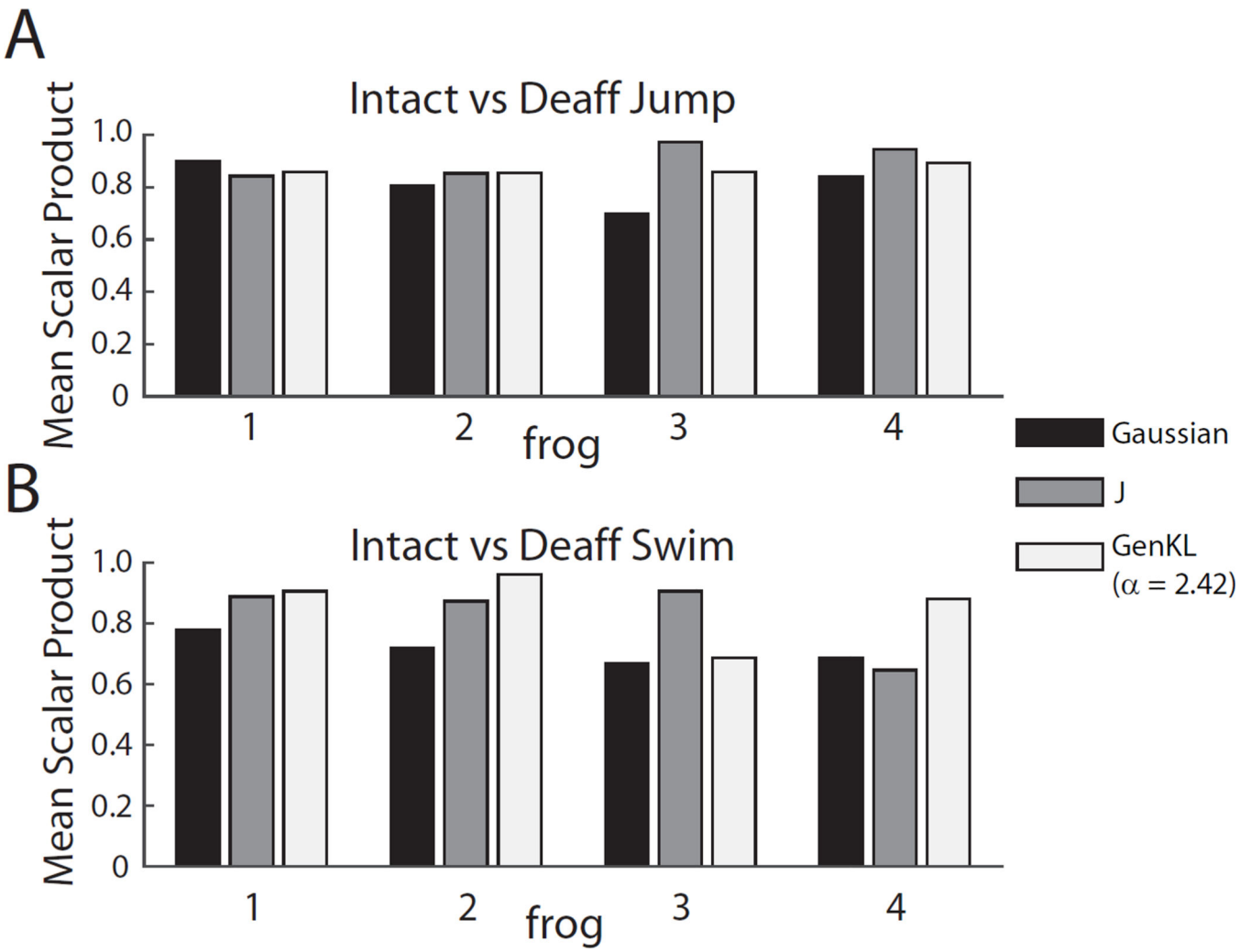

Figure 2.

The *GenKL* algorithm (α = 2.42) outperformed the Gaussian, and performed similarly well as the *J* algorithm. Performance of each algorithm was indicated by the similarity between the intact and deafferented muscle synergies, quantified by the best-match scalar product averaged across the synergy set of each frog (3 synergies for jump, and 4 synergies for swim). A, For jump, in all four frogs performance of the *GenKL* algorithm was comparable to that of the *J* in finding synergies that persisted after deafferentation, one of the best-performing algorithms in our previous study (Devarajan & Cheung, 2014). B, For swim, in three out of four frogs *GenKL* performed comparably well or even better than *J*.

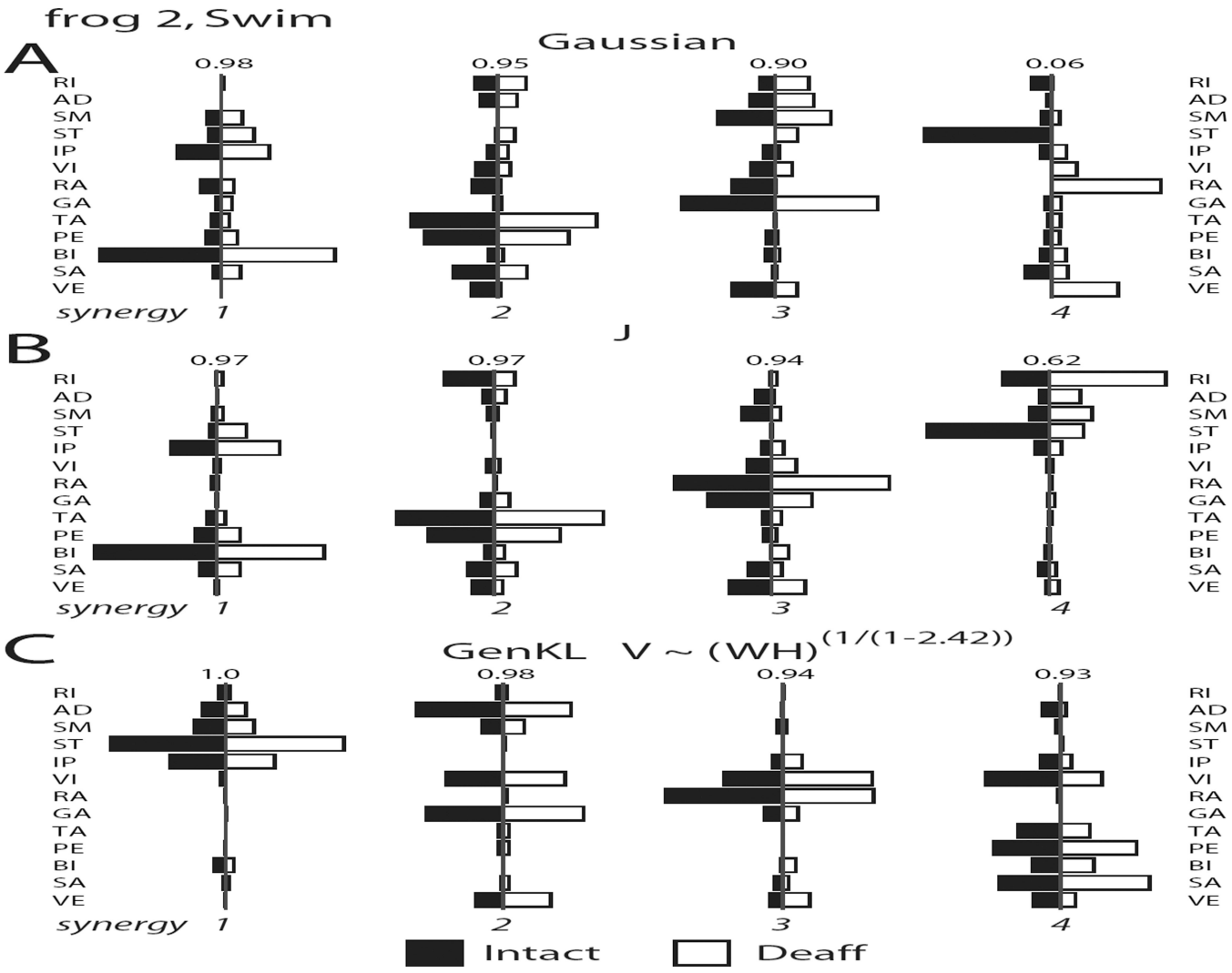

Figure 3.

Intact and deafferented muscle synergies of frog 2 for swimming extracted by the Gaussian, *J* and *GenKL* (α = 2.42) algorithms. Panels A (Gaussian) and B (*J*) are identical to Figure 5 of Devarajan & Cheung (2014). Overall, the *GenKL* algorithm (C) returned intact synergies (black bars) that were the most similar to the deafferented synergies (white bars). Abbreviations: RI, rectus internus; AD, adductor magnus; SM, semimembranosus; ST, semitendinosus; IP, iliopsoas; VI, vastus internus; RA, rectus anterior; GA, gastrocnemius; TA, tibialis anterior; PE, peroneus; BI, biceps femoris; SA, sartorius; VE, vastus externus.

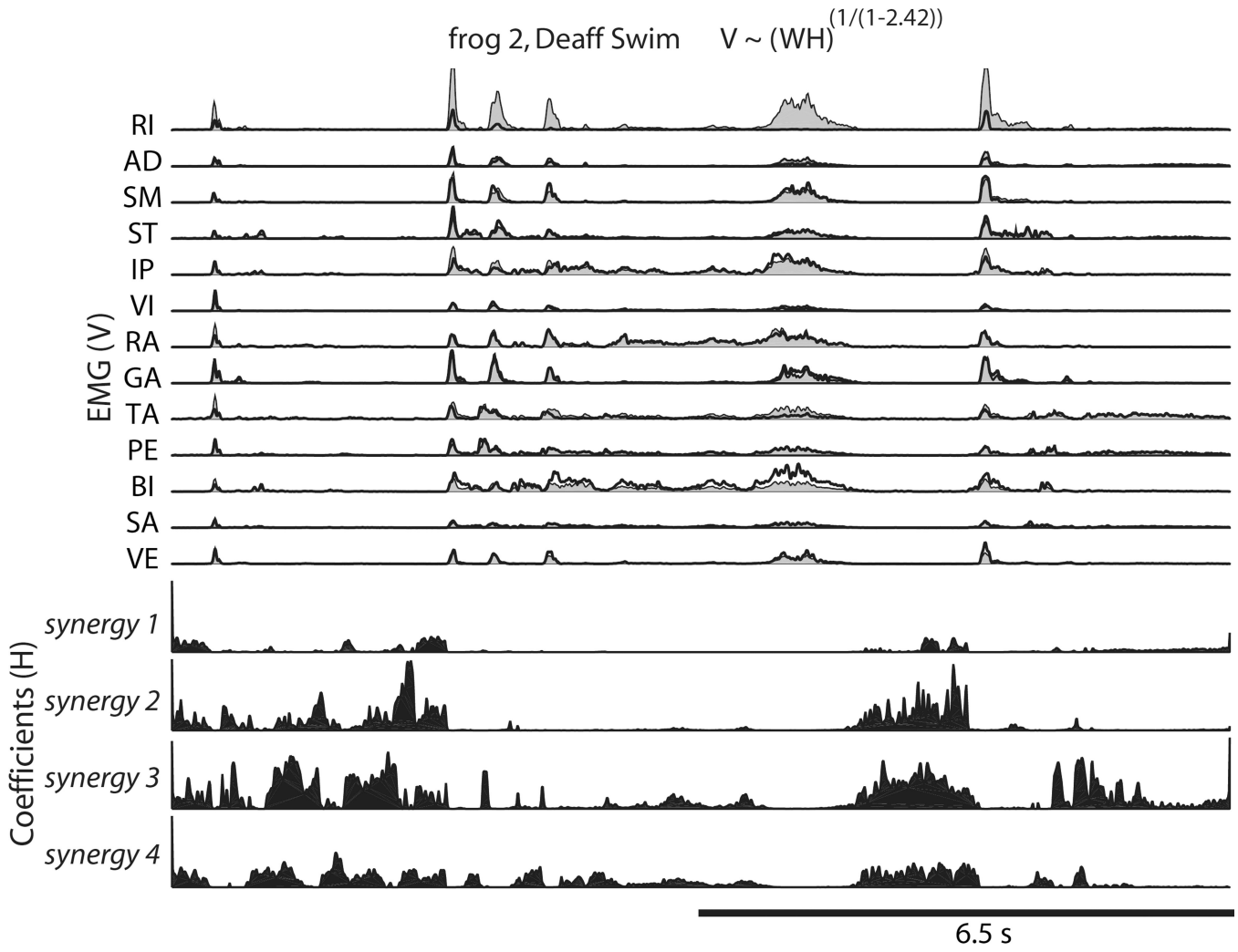

Figure 4.

Reconstruction of a swimming episode (frog 2) recorded after deafferentation by combining the *GenKL* muscle synergies and their time-varying coefficients via the inverse power link ($\alpha = 2.42$). The actual recorded EMGs are shown in thick black lines, and their reconstructions are shown as gray shadings. The coefficients for synergies 1 to 4, shown in the bottom half of the figure, correspond to the synergies 1 to 4 shown in Figure 3C. Notice that the coefficient of a synergy tends to be active when the EMGs of the muscles in the synergy are inactive. The synergies returned by the *GenKL* algorithm may contain information on a group of muscles are activated together, in a coordinated fashion, via synchronous disinhibition.

Table 1

Variance and link functions for various statistical models

Model	α	Variance $\Sigma_\alpha(\mu)$	Link $g(\mu)$	$\frac{d\theta}{d\eta}$
Gaussian	0	1	identity [*,1]	1
Poisson	1 [†]	μ	identity [1]	μ^{-1}
Compound Poisson	(1, 2)	μ^α	inverse power [*]	$1/(1-\alpha)$
Compound Poisson	(1, 2)	μ^α	identity [1]	$\mu^{-\alpha}$
Gamma	2 [†]	μ^2	inverse [*]	-1
Gamma	2 [†]	μ^2	identity [1]	μ^{-2}
Positive Stable	$(2, \infty)$	μ^α	inverse power [*]	$1/(1-\alpha)$
Positive Stable	$(2, \infty)$	μ^α	identity [1]	$\mu^{-\alpha}$
Inverse Gaussian	3	μ^3	inverse squared [*]	$-1/2$
Inverse Gaussian	3	μ^3	identity [1]	μ^{-3}

[*] $g(\mu) = \mu^{1-\alpha}$

[1] $g(\mu) = \mu$

[†] in the limit

Neural Comput. Author manuscript; available in PMC 2017 August 09.

Table 2

Comparison of algorithms

Model[a]	Divergence	Link	α	Deaff. Jump		Intact Jump		Deaff. Swim		Intact Swim	
				r	R^2	*r*	R^2	*r*	R^2	*r*	R^2
Gaussian	*GenKL*, β[b]	identity[†§]	0	2	79.06	2	79.00	2	61.26	1	50.27
Poisson	β[c]	identity[§]	1[*]	1	77.89	1	75.00	1	59.22	1	57.56
Compound Poisson	*GenKL*[f]	inverse power	1.5	1	82.64	1	80.03	1	61.93	1	59.37
Gamma	β[d]	identity[§]	2[*]	1	83.90	1	81.92	2	73.46	2	71.81
Gamma	*GenKL*[f]	inverse[†]	2[*]	1	83.90	1	81.92	2	74.26	2	72.01
Gamma	*dualKL*[e]	-	-	1	97.80	1	97.17	1	90.84	1	88.8
Gamma	*J*[e]	-	-	3	97.82	3	97.30	4	93.6	4	92.82
Inverse Gaussian	β[d]	identity[§]	3	11	99.07	11	99.15	12	99.36	11	99.04
Inverse Gaussian	*GenKL*[f]	inverse squared[†]	3	13	99.93	12	99.54	12	99.61	12	99.59
Inverse Gaussian	*dualKL*[e]	-	-	11	99.96	11	99.95	12	99.92	10	99.83
Positive Stable	*GenKL*[f]	inverse power[†]	2.25	2	88.30	2	87.24	3	81.59	3	80.68
Positive Stable	*GenKL*[f]	inverse power[†]	2.30	2	88.06	2	87.08	3	81.40	4	85.88
Positive Stable	*GenKL*[f]	inverse power[†]	2.34	3	91.62	2	86.93	4	86.18	4	85.75
Positive Stable	*GenKL*[f]	inverse power[†]	2.38	3	91.47	2	86.77	4	86.03	4	85.64
Positive Stable	*GenKL*[f]	inverse power[†]	2.42	3	91.32	3	89.78	4	85.85	4	85.51
Positive Stable	*GenKL*[f]	inverse power[†]	2.50	3	90.96	4	92.25	5	88.78	5	89.20

[†] inverse power

[*] in the limit

[a] All models are members of the exponential family; specific models are identified by their standard names.

[b] For the Gaussian model, this is commonly referred to as Euclidean distance. EM algorithm proposed in Lee & Seung (2001).

[c] EM algorithm proposed in Lee & Seung (2001).

[d] MM algorithm proposed in Févotte & Idier (2011) based on β-divergence ([§] obtained using $g(\mu) = \mu$ in eqn. (6)). Results are similar to those from heuristic algorithms of Cheung & Tresch (2005) and Cichocki et al. (2006) (data not shown).

[e] EM algorithm proposed in Devarajan & Cheung (2014) based on dual *KL* or *J*-divergence

[f] Proposed family of algorithms based on *GenKL* divergence in eqn. (10).